Forthcoming, Production and Operations Management

Analyzing the Proposed Reconfiguration of Accident-and-Emergency Facilities in England

Houyuan Jiang

h.jiang@jbs.cam.ac.uk

Judge Business School

University of Cambridge

Trumpington Street

Cambridge CB2 1AG, UK

+44 (0)1223 765466

Manmohan S. Sodhi [1]

m.sodhi@city.ac.uk

Cass Business School

City, University of London

106 Bunhill Row

London EC1Y 8TZ, UK

+44 (0)20 7040 0276

Abstract

The Keogh Report of 2013 proposed a major reconfiguration of the accident and emergency (A&E) system under National Health Service (NHS) England to improve service. The proposed reconfiguration includes centralized facilities with multiple specialties as well as small local minor-injury facilities. We use stylized queuing models to analyze cost and service implications of the proposed reconfiguration. We find that increasing numbers of specialty patients that require admission to hospital makes splitting off specialty A&Es from general ones more attractive. The same applies for patients with minor injuries. Our work generally supports the reconfiguration recommended in

[1] Corresponding author

the Keogh report but with some fine-tuning: For instance, a merger of A&Es (pooling) does not always make sense even though it increases patient numbers when the patients in the two A&Es are of different types. We provide simple quantitative rules to indicate whether the proposed reconfiguration could lower costs in any particular region of the country. The results here are consistent with some NHS England providers attempting specialty A&Es for geriatric patients and mobile drunkenness treatment centers on weekends. Our rules and approach can be useful for identifying candidate reconfiguration opportunities not only for NHS England but also for any other context where pooling and arrival heterogeneity are important considerations.

Keywords: *Healthcare policy, accident-and-emergency service, queuing models, pooling and splitting, merger*

History: Received: September 2018, Accepted: February 2019 by Michael Pinedo, after two revisions.

1. Introduction

In accident-and-emergency (A&E) centers in England, attendance grew 67% from around 14 million per year prior to 2003 to 24 million in 2017–2018 (NHS England 2018), resulting in a sharp deterioration in service. Many hospital providers are in violation of the policy-mandated service-level of no more than 5% of patients having to wait longer than four hours from the time of entering the A&E to being formally released. To improve matters, Professor Sir Bruce Edward Keogh, National Medical Director (2007-13), recommended fundamental reconfiguration of the A&E system with the creation of mega-facilities in urban areas with multiple specialties as well as smaller local sites for minor injuries (NHS 2013). Such a reconfiguration could be realized by, for instance, merging existing generalist facilities across hospitals and then hiving off multiple specialty clinics to treat specific categories of patients. We analyze the cost implications of such reconfiguration, and in doing so, provide simple queuing-theory-based rules for such evaluation.

NHS England has three types of A&E facilities: *Type 1* for all emergency patients, *Type 2*, with a single specialty service such as ophthalmology-only, dentistry-only, or trauma-related emergency services, and *Type 3*, for minor injuries only. A hospital may have any subset of these facilities and may have more than one Type-2 A&Es for different specialties. The Keogh report recommendations included a two-tier system in urban areas: small neighborhood-level Type-3 A&Es for minor injuries at facilities including pharmacists and "mega" Type-1 facilities supported by multiple Type-2 A&Es at a few large centralized hospitals. Such a configuration could be achieved in urban areas by creating new Type-3 A&Es, say, at pharmacists, by merging Type-1 facilities making them much bigger, and by splitting off Type-2 services from the existing or merged Type-1 A&Es. The recommendations also included "developing models and tools to improve...the management of capacity", which this paper seeks to do.

We develop stylized queuing models to identify the necessary and sufficient conditions to lower costs while maintaining the target service level, noting that these simple models are robust against more realistic assumptions that would be intractable analytically. *First*, we obtain the necessary

and sufficient conditions for splitting off a specialty Type-2 A&E from a general Type-1 A&E, as for instance, in creating a specialty A&E for geriatric patients. *Second*, we do the same for splitting off a minor-injury Type-3 A&E from a Type-1 A&E, as for instance, using "booze buses" in city centers on weekends to provide supplementary A&E services for over 2 million alcohol-related emergencies ("Drunks should be treated in 'booze buses' to ease A&E overcrowding, nurses say", Daily Telegraph, 16 June, 2014). *Lastly*, we analyze the potential merger of two hospitals for optimal reconfiguration of their A&Es.

Our contribution to the literature is demonstrating the application of queuing models to healthcare policy regarding a nationwide system of A&E facilities. We seek to contribute (a) to the healthcare operations literature on policy by analyzing the Keogh recommendations, and (b) to the queuing theory literature – including Cachon and Terwiesch (2009), van Dijk (2008), Mandelbaum and Reiman (1998) and Song et al. (2015) – with a real-world application of pooling (or splitting), given heterogenous arrivals. The setting we describe has alternative A&E configurations that have not received enough attention in the literature thus far although Cawson et al. (2012) and Green (2012) have highlighted the creation of specialty units and reconfiguration of service units as important topics. Moreover, A&E departments have not been considered from a policy perspective at a system-wide level in the operations literature to the best of our knowledge.

Our indicative results have at least two managerial implications. *First*, while Keogh recommendations generally make sense from a cost perspective, equally, there are contexts where mergers to produce mega-facilities may not make sense. *Second*, our work has yielded simply-to-apply rules for policymakers to do a first-pass evaluation of any reconfiguration whether or not motivated by the Keogh recommendations. These rules can indicate or rule out candidate reconfigurations for a pair of hospitals or even for the A&E system as a whole for a region. For instance, these rules indicate when it makes sense to create Type-2 A&Es for the elderly – a growing percentage of the population – or to create Type-3 A&Es to respond to minor injuries, including those related to weekend drunkenness in town centers, based on arrival rates. Similarly, these rules indicate how

an increase in specialty-patient arrivals or in their admission rates to hospital makes splitting off specialty Type-2 A&Es more attractive. Applying these rules to a merger of two hospitals' A&Es also indicate when it does not make sense to merge.

Section 2 looks at the pertinent queuing literature. Section 3 provides the real-world context for this work and modeling preliminaries. In Section 4, we develop conditions under which splitting off a specialty Type-2 A&E from a general Type-1 A&E is cost-effective while Section 5 does so for splitting off a Type-3 A&E for minor injuries from a general Type-1 A&E. Section 6 develops appropriate conditions for reconfigurations of a pair of A&E facilities targeted for merger. Section 7 concludes with ideas for further research.

2. Literature

Many healthcare systems have been analyzed using queueing theory; for example, intensive care units (ICUs) (Chan et al. 2012 and Chan et al. 2014), general practice (Green and Savin 2008) and outpatient services (Jiang et. al 2012). Regarding NHS England specifically, Mayhew and Smith (2008) use queuing theory to analyze the four-hour completion target for A&E departments. Armony et al. (2011) investigate an A&Es as just one part of a hospital as a queueing network. Saghafian et al. (2015) and Saghafian et al. (2012) use queueing models – along with Markov decision processes (MDPs) and hospital data – to study issues related to patient flow, patient streaming, triage, and patient sequencing in A&Es. Other researchers have used mathematical programming, optimization and simulation to analyze A&E performance; see review by Saghafian et al. (2015).

Our study pertains to the benefits of pooling in simple and network queues because we investigate whether or not it is more cost efficient to create specialty A&Es from an existing A&E or to merge and/or reconfigure a pair of A&Es. The literature on pooling suggests that while pooling lowers costs when patients are homogenous, doing so may increase costs when patients are heterogenous as doctors would need to have a broader range of skills, setup would be increased, and variability of the service processes would increase. In this thread, Smith and Whitt (1981) show

that operating a single queueing system with $n_1 + n_2$ servers is at least as effective as operating two independent queueing systems with n_1 and n_2 servers respectively, when customer inter-arrival and service times are identically distributed for all facilities. Benjaafar (1995) extends this work for n independent queueing systems by providing performance bounds on the effectiveness of several pooling scenarios and by offering capacity and utilization tradeoffs between independent and pooled queueing systems. Mandelbaum and Reiman (1998) consider a particular queueing network where the tasks at all nodes of the queueing network are processed by a single super server, and compare independent and pooled systems under assumptions for traffic intensity and task variability. Andradottir et al. (2017) study the effects of resource pooling in the presence of failures. They show that while pooling queues is always beneficial, pooling servers and queues increases risk although it does improve efficiency. Cachon and Terwiesch (2009) summarize benefits and drawbacks of pooling; see also van Dijk (2008), Mandelbaum and Reiman (1998), and Song et al. (2015).

There is a gap in this literature regarding the application of queuing models to systemwide multiple A&Es. Our paper contributes in narrowing this gap by providing and analyzing a particular real-world setting. Moreover, in our setting, heterogeneity stems from different treatment of patients in a second stage in the queuing system. Our two-station tandem queue in the pooled system is unlike Smith and Whitt (1981) or Benjaafar (1995) who have only one station in independent and pooling queueing systems; the two-stage queue is used by Conroy et al. (2014) and Wright et al. (2013). Our work is also different from Mandelbaum and Reiman (1998) because, in the pooled system of our queueing network, tasks in different nodes are processed by different servers whereas in the pooled system of Mandelbaum and Reiman (1998), the tasks at different nodes are sequentially processed by a single super-server. We do not consider service failures like Andradottir et al. (2017), but such failures only benefit pooling. Therefore, the pooling effect in our setting is lower than that in the queueing systems in Smith and Whitt (1981) and Benjaafar (1995). The implication is that our analysis is not a straightforward application of models from the existing literature to compare pooled and independent queuing systems.

3. Background and Modeling Preliminaries

Healthcare in the UK is devolved to the four constituent countries: England, Northern Ireland, Scotland, and Wales. NHS England, as a public body responsible for health services for all residents of England, in turn commissions *provider organizations* that include NHS Trusts, NHS Foundation Trusts and private or independent sector organizations (ISO). Each provider organization manages one or more hospitals that may offer A&E services or ambulance service. Of the 247 providers in 2015-16, 138 offered Type-1 A&Es, 31 offered Type-2 A&Es, 171 offered Type-3 A&Es, and 10 offered no A&E (Table 1; NHS England (2017b)).

No. provider organizations	16	10	60	1	52	4	94	10	247
Type-1 services	✓	✓	✓		✓				138
Type-2 services	✓	✓		✓		✓			31
Type-3 services	✓		✓	✓			✓		171

Table 1 Number of provider organizations offering A&E services in NHS England, March 2016.

Individual hospitals, not just their providers, may also offer more than one type of A&E services (NHS England 2017a). Queens Hospital Romford operates one Type-1 A&E and several Type-2 A&Es, which are for trauma, hyper-acute stroke, maternity, renal and neurosciences patients respectively. In contrast, UCL Hospitals London offers only one Type-1 A&E. Moorfields Eye Hospital in London has only a Type-2 A&E for ophthalmology but no Type-1 service. As, Table 1 shows, as many as 94 providers only have Type-3 A&Es to deal with "minor injuries" that only have nurses (no doctors) so the facilities are less expensive to operate than Type-1 A&Es.

Patient flow. In case of an accident or other emergency, a patient can go to any A&E in any hospital anywhere in the country nearest to them or be taken there by ambulance. (The ambulance can take the patient directly to a Type-2 A&E, if applicable.) Upon arrival, the patient is registered and the time noted. Shortly thereafter, a nurse performs triage to assess the urgency and severity of the patient's condition to assign priority. (If necessary, and if an appropriate Type-2 A&E is

available on-site, the patient is directed there.) Next, the patient is seen by a doctor or nurse. Tests may be carried out as necessary with results seen by the treating doctor. Eventually, the patient is released from the A&E (time noted) in one of three ways: (a) discharge, (b) release for admission to hospital and is placed in a temporary ward or on a trolley to taken to the appropriate ward (the time between the decision to admit and that when admission actually takes place is also monitored), or (c) transfer to some other A&E facility.

Waiting time and service policy. The difference between the release timestamp and the arrival timestamp is the *episode time* – also referred to as *waiting time* or *sojourn time* in the queuing literature – in the A&E system, which is an important performance measure for providers and of NHS England as a whole. The provider organizations' service level requirement requires that no more than 5% of all patients over a measured period (a month) have waiting time exceeding *four* hours. All providers are required to report this statistic, aggregated across the hospitals they manage, as part of Hospital Episode Statistics (HES). Providers get paid based on arrivals and treatments but are liable for a penalty of £120 per patient for all patients, whose waiting time exceeds four hours above the 5% threshold (Department of Health 2016; Propper et al. 2008).

3.1. Queuing model

Although there are several performance measures such as the average queuing length, the average waiting time, the average number of customers in the system, and the average sojourn time, we focus on the tail probability for the sojourn time because it is a crucial performance measure with service level specified as $P(W > T) \leq \alpha$. For NHS England, $T = 4$ hours and $\alpha = 5\%$ requiring that more than 95% of all patients arriving in an A&E must be released – discharged, released for admission to hospital, or transferred to some other facility – within four hours. In other words, less than 5% of patients should have to wait longer than 4 hours. If the A&E were an $M/M/1$ system, we would have $P(W > T) = e^{-(\nu-\lambda)T}$ for arrival rate λ and service rate ν. The service level requirement would then be equivalent to having a service rate

$$\nu \geq \lambda + \frac{1}{T} \ln\left(\frac{1}{\alpha}\right), \tag{1}$$

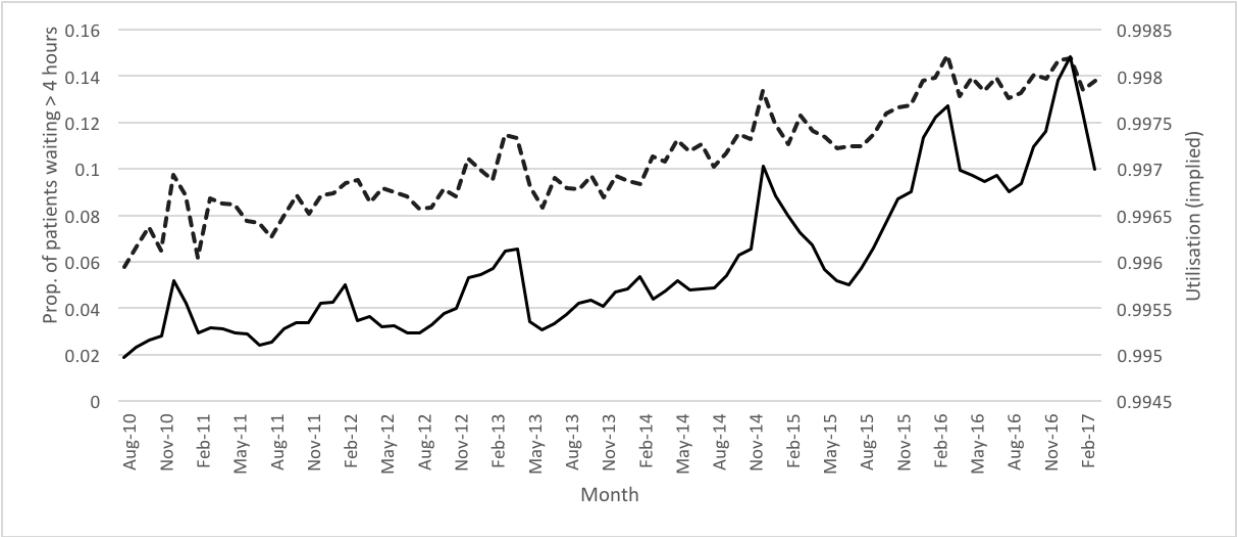

Figure 1 **Proportion of patients waiting in excess of four hours by month (solid line, left axis) with implied utilization of A&E services in NHS England as a whole (dashed line, right axis). NHS England, May 2017.**

thus providing the *minimum acceptable service rate* for this single-server A&E department. In reality, an A&E is more complex than a $G/G/s$ queue. However, an approximation helps us to model A&E performance for analytical tractability under heavy traffic when utilization is close to 100% as is the case here (Figure 1).

The heavy traffic assumption allows us to assume, with justification from the literature, that the tail probability for waiting times queue can be approximated by an exponential function for many queuing systems under heavy traffic. This fact is especially useful here because service level specification in NHS England is also in terms of tail probability. As such, throughout the paper, our basis for analysis is

ASSUMPTION 1. *The A&E operation has Poisson arrivals at rate λ that creates heavy traffic in that, if μ is overall service rate of the system, then the utilization $\lambda/\mu \to 1$. The tail probability of the waiting time in the system, W, for large T is characterized by $P(W > T) \approx e^{-\kappa(\mu-\lambda)T}$, where and the parameter κ is independent of the arrival rate.*

There is a substantial body of literature on the exponential approximation under heavy traffic and large T for different queuing systems including the $M/G/s$ queue. Abate et. al (1995) analyze

exponential approximations and Allon and Federgruen (2008) point out that such an approximation becomes exact for the $G/M/s$ queue. Specifically for an $M/G/s$ system under heavy traffic, exponential approximation for the tail probability of the sojourn time, W, for large T is given by

$$P(W > T) \approx \varphi e^{-\eta T}, \tag{2}$$

where φ and η depend on the characteristics of the queueing system. In **Appendix A**, we show that under Assumption 1, φ and η can be approximated as constants:

$$\varphi \approx 1, \quad \eta \approx \frac{2}{1+\tau_s^2}(s\nu - \lambda) \tag{3}$$

where λ is the arrival rate per hour, ν the service rate per hour for one server, and τ_s is the *coefficient of variation* of the service time. Following (2) and (3), we approximate $P(W > T)$ by $e^{-\kappa(s\nu-\lambda)T}$, where $\kappa = 2/(1+\tau_s^2)$. For the $M/M/1$ queue, the approximation is exact as we saw earlier. Whitt (1993) reviews prior results and provides new approximations for $G/G/s$ queue with an exponentially decaying length of queue under heavy traffic (p.121).

Aggregate service rate approximation. Taking $\mu = s\nu$, under Assumption 1, the service level requirement $P(W > T) \leq \alpha$ implies that we require an aggregated service rate μ given approximately by

$$\mu \geq \lambda + \frac{1}{\kappa T} \ln\left(\frac{1}{\alpha}\right). \tag{4}$$

to meet service level requirements. We use this approximation throughout the paper and refer to μ as the *aggregate service rate* for the queuing system. When ν is a fixed parameter, calculating the required value of μ is equivalent to determining the number of servers s in the queue. Equation (4) indicates that the minimum aggregate service rate in the system must exceed the arrival rate plus an additional capacity to meet the service requirements characterized by the waiting time target T and service level α.

Our choice of any queuing system satisfying Assumption 1 extends some of the existing literature using $M/M/1$ queues (cf. Green and Savin 2008 and Jiang et. al 2012). The assumption allows

us to assume, for instance, the $M/G/s$ system to represent the patient flow from A&Es to hospital wards. This is still a simplification but it makes the modeling tractable for generating managerial insight for answering policy-level questions such as those addressed in NHS (2013).

4. Case 1: Splitting off Type-2 A&Es from Type-1 A&Es

Many A&Es in NHS England are of Type 1 and treat all patients. A possible reconfiguration would be to split off a Type-2 A&E for *specialty* patients who can immediately be categorized for a specialty facility requiring, for instance, opthalmology, trauma or cardiac treatment. Other patients would remain in the *general* pool for the Type-1 A&E. Taking this idea further, multiple Type-2 A&Es can be created for different categories of specialty patients in the same hospital.

Consider **pooled** and **split** systems as in Figure 2 panels (a) and (b) respectively, along with some notation we explain next.

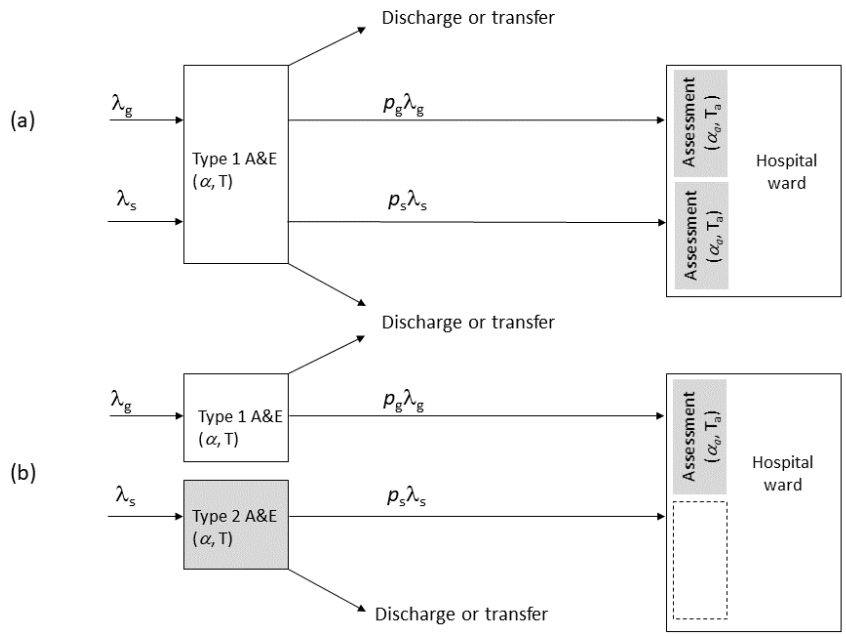

Figure 2 An illustration of A&E configurations. (a) a pooled system. (b) a split system.

Pooled system. For a pooled or Type-1 A&E, general as well as specialty patients arrive at the A&E at the rate λ_g and λ_s respectively. They receive services provided by generalist doctors,

not specialists. A proportion of the patients who visit the A&E have to be admitted to hospital for further care from specialist doctors, with the remaining are either discharged or transferred to other hospitals. The proportion of general patients and specialty patients released for admission to hospital wards are p_g and p_s respectively. Any procedures or diagnoses that these released specialty patients receive in A&E from generalist doctors will generally be repeated by specialty doctors in hospital wards (Geddes 2013). Waiting time to get admitted (and get these preliminary services in hospital wards) is specified by T_s with α ensuring a service level for the waiting time for formal admission to hospital ward, with T_s specified at two thresholds, 4 hours and 12 hours. Thus, we assume that admitted patients receive the same services at either stage – by generalists in the A&E and by specialists in hospital wards. We treat these two stages individually *as if* the output of the first stage is Poisson, noting that Burke (1956) has shown that the output is Poisson for an $M/M/s$ queue as has Mirasol (1963) for an $M/G/\infty$ system. It follows from (4) that:

$$\mu_{g1} \geq \lambda_g + \lambda_s + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right), \tag{5}$$

where μ_{g1} is the service rate (capacity) for meeting the required service level specified by T_g and α for the queue in the first stage of the pooled system. Similarly, the service-level requirement in the hospital ward for the preliminary services for specialty patients can be met if:

$$\mu_{s1} \geq p_s\lambda_s + \frac{1}{\kappa T_s} \ln\left(\frac{1}{\alpha}\right), \tag{6}$$

where μ_{s1} is the service rate (capacity) for meeting the required service level specified by T_s and α for the specialty-patient queue in the second stage of the pooled system. As for costs of the pooled system, let c_g and c_s be the unit costs per service rate for generalists and specialists, respectively. Then the *total hourly cost* for the pooled system is $c_g\mu_{g1} + c_s\mu_{s1}$. Thus, (5) and (6) show that the *minimum total cost for the pooled system* per hour is

$$\pi_1 = c_g\left(\lambda_g + \lambda_s + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right) + c_s\left(p_s\lambda_s + \frac{1}{\kappa T_s} \ln\left(\frac{1}{\alpha}\right)\right). \tag{7}$$

Split system. By contrast, for a split system, where a Type-2 A&E has been split off from the existing Type-1 A&E (panel (b) of Figure 2), general patients and specialty patients enter the A&E

through separate queues for their respective Type-1 and Type-2 A&Es. All general patients are treated in the A&E by generalists in one $M/G/s$ queue and specialty patients (identified through triage) are treated by specialists in another $M/G/s$ queue. As with the pooled system, patients are discharged, released for admission to hospital, or transferred to other hospitals. Admitted patients will receive further services and treatments in hospital wards. but duplication of preliminary assessment is avoided in the split system. Bringing specialists into A&Es can also improve service quality for specialty patients (Conroy et al. 2014) but specialists doctors are more expensive than generalist doctors. We use equation (4) to obtain service rates μ_{g2} and μ_{s2} such that $\mu_{g2} \geq \lambda_g + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)$ and $\mu_{s2} \geq \lambda_s + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)$. Then, the *minimum total cost for the split system* is

$$\pi_2 = c_g \left(\lambda_g + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right) + c_s \left(\lambda_s + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right). \tag{8}$$

4.1. Comparison between Pooled and Split System

We now compare the minimum costs of the two configurations even though the relationship between minimum costs of these configurations is not the same as the relationship between their actual (or projected) costs. This is partly because we look at this issue from a policy perspective. So, we do not include all costs especially if these appear to be equal for the two configurations. This would be the case with triage costs when a nurse has to categorize a patient – in the configurations we compare, there is one triage in either configuration, so we ignore this cost. Moreover, to compare two configurations, we necessarily compare their optimized minimum costs because if one configuration is assumed to be inefficient, its cost can always be lowered by making it run efficiently and eliminating service level violations.

Therefore, splitting off a specialized Type-2 A&E from a pooled Type-1 A&E makes sense from a cost perspective if and only if $\pi_1 \geq \pi_2$, i.e.,

$$c_g \left(\lambda_g + \lambda_s + \tfrac{1}{\kappa T_g} \ln\left(\tfrac{1}{\alpha}\right)\right) + c_s \left(p_s \lambda_s + \tfrac{1}{\kappa T_s} \ln\left(\tfrac{1}{\alpha}\right)\right) \geq c_g \left(\lambda_g + \tfrac{1}{\kappa T_g} \ln\left(\tfrac{1}{\alpha}\right)\right) + c_s \left(\lambda_s + \tfrac{1}{\kappa T_g} \ln\left(\tfrac{1}{\alpha}\right)\right). \tag{9}$$

We do not account for servicing general patients in hospital wards because the costs in pooled and split systems are identical in both systems, and therefore cancel out in (9). Rearranging terms in (9), we obtain

PROPOSITION 1. *A split system is less costly than a pooled system if and only if* $p_s \geq \frac{1}{\lambda_s}\left(\frac{1}{\kappa T_g} - \frac{1}{\kappa T_s}\right)\ln\left(\frac{1}{\alpha}\right) + 1 - \frac{c_g}{c_s}$.

Proposition 1 implies that splitting off a separate Type-2 A&E becomes more attractive if there are many specialty patients or if specialized capacity is cheap. For the case $T_s = T_g = 4$ hours and specialty cost $c_s = \frac{3}{2}c_g$, a split is indicated if $p_s > 1/3$. For the higher threshold of $T_s = 12$ hours, p_s will need to be higher for the split to be cost-effective although this increase can be mitigated by an increasing arrival rate λ_s of specialty patients. Larger values for p_s or λ_s makes the second-stage service for specialty patients in a pooled (Type-1) system more costly. Although T_g is fixed by government policy, a larger value would favor pooling, which is an idea floated by the Minister for Health (Guardian, 9 Jan 2017).

Finally, the arrival rate for general patients λ_g is irrelevant to the choice of a pooled or split system. *This means we can apply the rule in the proposition sequentially to split off any number of non-overlapping specialty Type-2 A&E.* This allows us to evaluate the creation of mega-A&E centers in urban areas in England with multiple Type-2 specialty A&Es (one for each specialty) as recommended by the Keogh report.

Proposition 1 implies that the benefit of pooling is reduced and may even become negative when customer heterogeneity increases. This is in line with van Dijk et al. (2008) in that there is no single answer to the question of whether service capacity should be pooled or not. Green (2012) has used intensive care units (ICU) to illustrate the advantages and disadvantages of specialization and pooling and highlights the economies of scale generated by combining two or more ICUs. Finally, while the proposition focuses on only one benefit of a split system over a pooled one, there are other benefits – dedicated units reduce variability in treatment and length of stay, and enable better coordination with other hospital units (Song et al. 2015).

4.2. Implications for specialty A&Es for geriatric patients

Elderly patients visiting A&Es in England have increased steadily in recent years both in absolute numbers and as a percentage of total attendances (Age UK (2017b)) which was already about

20% in 2013 (Geddes 2013). Moreover, total costs for geriatric patients are disproportionate: over 80% of admitted patients whose length of stay longer than two weeks are those who are over 65 years old (Wright et al. 2013). Some of these admissions could be avoided with A&Es devoted to geriatric patients (Geddes 2013). In addition, the proportion p_s of elderly patients being admitted for further care is quite high at nearly half (47%), varying from about 40% for those over 65 to over 60% for those over 80. Therefore there is need to consider splitting geriatric patients off from Type-1 A&Es.

The rule from Proposition 1, using $T_g = 4$ hours, $\alpha = 5\%$ (so $\ln(1/\alpha) \approx 3$) and $T_s = 12$ hours, approximates to $p_s - \frac{1}{2\lambda_s} \geq 1 - \frac{c_g}{c_s}$. Assuming exponential service distribution, $\tau_s = 1$ and consequently $\kappa = 1$ as well. We use $\lambda_s = 0.2\lambda$ (the total arrival rate $\lambda = \lambda_s + \lambda_g$) and assume $c_s = 1.5c_g$, i.e., specialists are 50% more expensive than generalists, to compare with $p_s = 47\%$ to check if any of the 52 providers currently offered Type-1-only A&Es in 2016 should consider offering Type-2 A&Es for geriatric patients. Many of these 52 providers satisfy the above inequality (see the light grey bars for each provider relative to the horizontal line in Figure 3). However, if we assume $c_s = 1.2c_g$, i.e., the cost of specialists were only 20% more than that of generalists, realized by using a higher proportion of nurses, then nearly all of the 52 providers can justify having geriatric Type-2 A&Es (see the dark grey bars of Figure 3).

This cost-based justification is only an indication and further analysis is necessary. Elderly patients have high readmission rates compared to the general population. This means higher costs if some elderly patients are incorrectly discharged in a Type-1 A&E or do not get the right treatment in the general hospital; this is why p_s for geriatric patients can be high to begin with, resulting in admitting patients who would have been discharged in a dedicated Type-2 A&E. Thus, a separate Type-2 A&E dedicated to elderly patients can lower costs for providers by reducing costs of admission such patients to hospital.

The literature provides empirical evidence for cost reduction and improved care. Wright et al. (2013) report that in September 2010, the Royal Free Hospital and Haverstock Healthcare Ltd, a

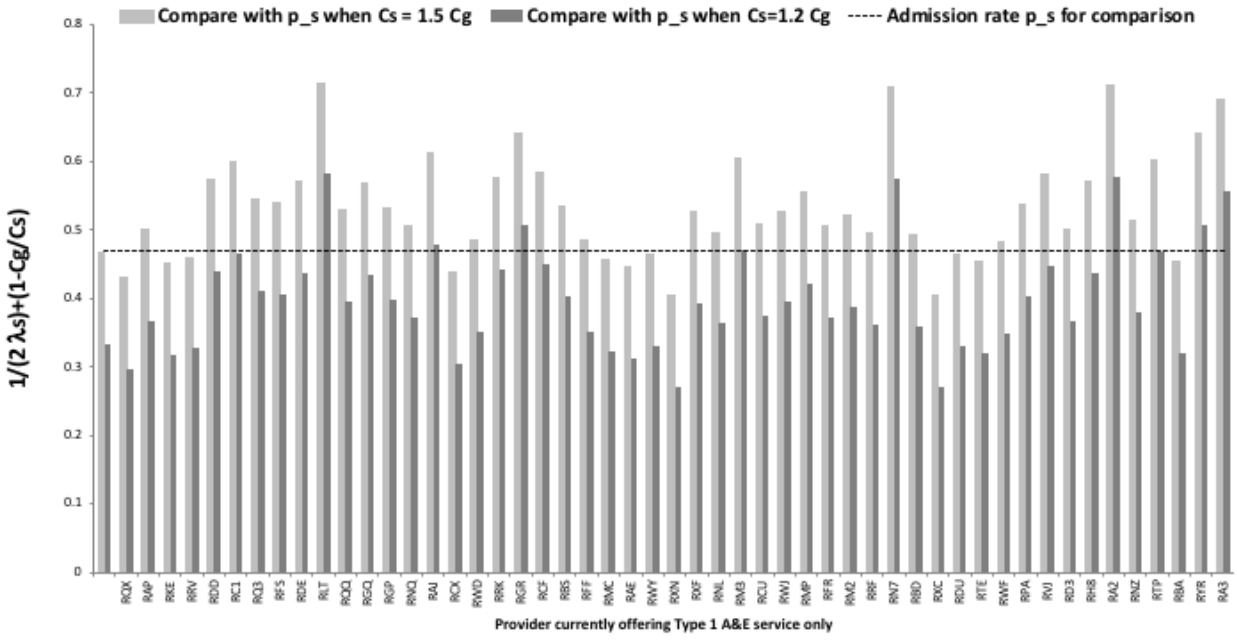

Figure 3 **Identifying 52 Type-1-only provider organizations for splitting off Type-2 A&Es for geriatric patients, assuming $c_s = 1.5c_g$ (light grey bars) or $c_s = 1.2c_g$ (dark grey bars).** *Source*: **NHS data, March 2016.**

general practitioner provider organization, introduced an admission-avoidance system for patients aged 70+, called the Triage and Rapid Elderly Assessment Team (TREAT). A study reported that TREAT reduced avoidable emergency geriatric admissions to the hospital and, in addition, shortened the length of stay in the hospital for all geriatric patients who were admitted. Conroy et al. (2014) report findings on a similarly motivated Comprehensive Geriatric Assessment (CGA) team formed after the merger of two acute medical services (Leicester Royal Infirmary and Leicester General Hospital) that resulted in improved discharge rates as well as reduced readmission rates for older patients after being discharged from the hospital. Geddes (2013) reports a similar intervention – with potential savings of £3m/year in hospital costs – in the North General Hospital in Sheffield without using extra staff other than a staff nurse, although specialist doctors (geriatricians) had to "adjust working hours so they were on call in the evenings and at weekends". In light of such evidence and our calculations from Proposition 1, there is a strong case for geriatric Type-2 A&Es systemwide.

5. Case 2: Splitting Type-3 A&Es from Type-1 A&Es

Consider general patients without minor injuries arriving at a rate of λ_g and those with minor injuries arriving at a rate of λ_m. For the latter, the pooled system is simpler because there is no second stage while the split system remains the same as that investigated in the previous section. Let T_g be the waiting time targets for general patients in either system. Although the waiting time requirement in Type-3 A&Es is currently the same as in a Type-1 A&E (T_g, α), it is worthwhile introducing a separate waiting time target, T_m, for the split A&E with a different service requirement α_m. This is based on the idea of lowering service levels for those with minor injuries floated by the UK Minister of Health (Guardian, Jan 9, 2017). Unit costs for serving general and minor injuries patients are respectively c_g and c_m, with $c_m < c_g$ as Type-3 A&Es are staffed by nurses rather than doctors.

Repeating the analysis in the previous section adapting (9) without the second stage on the left hand side and noting the potentially different service levels for the Type-3 A&E for minor injuries, we obtain:

PROPOSITION 2. *Splitting off a dedicated Type-3 facility for minor injuries with $c_m < c_g$ from a pooled Type-1 facility lowers the service cost if and only if*

$$\lambda_m \geq \left(\frac{c_m}{c_g - c_m}\right) \frac{1}{\kappa T_m} \ln\left(\frac{1}{\alpha_m}\right).$$

The above proposition provides a simple rule to aid decision-making just like Proposition 1 does. The main implication is that the lower the service-level requirements on Type 3 or the cheaper the capacity for Type 3, the better it is to split. The case for splitting is stronger in urban areas because of large λ_m. Indeed, if λ_m were large enough, splitting off multiple Type-3 A&Es from the same Type-1 facility may be justified. This supports the Keogh recommendation for multiple distributed Type-3 A&Es, including services being provided by, say, the local pharmacy. The proposition also shows that the case for splitting becomes stronger if the government were to weaken service levels for minor-injury patients, by increasing T_m or α_m, currently the same as with Type-1 A&Es (T_g

and α). Given an arrival rate of minor-injury patients of λ_m, a separate Type-3 A&E becomes more attractive with increasing values of c_g and less attractive with increasing value of c_m. Finally, as with specialty patients in the previous section, justification for splitting is independent of the arrival rate of general patients, λ_g. In similar vein, Mayhew and Smith (2008) advocate separating patients who require 'short' (minor) treatments from those who need 'long' (major) treatments. Cooke et al. (2002) provide empirical support for reduction in waiting time by using a variant of a Type-3 A&E within a Type-1 A&E in a UK hospital.

Unlike geriatric patients, minor-injury patients require triage for classification into the 'minor-injury' category. Ieraci et al. (2008) state that patient complexity must be factored into triage and streaming, while Saghafian et al. (2014) show that complexity-based triage, as in this case, lowers the risk of adverse patient events as well as the average length of stay.

5.1. Implications for providers

The Keogh report envisions more Type-3 A&Es because patients with minor injuries comprise a high proportion – maybe as much as 60% – of all A&E arrivals. The NHS has noted that 57.7% of all the 19m patients who visited A&Es in 2014-15 were discharged with only a GP follow-up or with no follow-up required. This is why Type-3 facilities are common with 171 of the 247 provider organizations in NHS England offering Type-3 A&Es in 2017, either in conjunction with a Type-1 or Type-2 A&E, or on their own. In 2016-17, nearly a third (32%) of all attendances were to Type 3 A&Es, in addition to patients with minor injuries who visited Type-1 A&Es.

Proposition 2 provides the following rule for splitting a Type-3 A&E from a Type-1 service: $(c_g/c_m) \geq 1 + \ln\left(\frac{1}{\alpha}\right)/(\kappa T_m \lambda_m)$. To apply the calculation to the 52 A&Es with Type-1 service only, take $T_m = T_g = 4$ hours, $\tau_s = 1$ and $\alpha = 0.05$. Assuming 40% of all patients at each of these providers have minor injuries only, we find this rule is satisfied for all but four of the 52 Type-1 A&Es, indicating that it is attractive to split off Type-3 A&Es (light grey bars in Figure 4). Of course, if the percentage were higher, say, 60%, the case becomes even stronger (dark grey bars in Figure 4). Thus, providers could lower costs by offering more and separate Type-3 A&Es as suggested by the Keogh report.

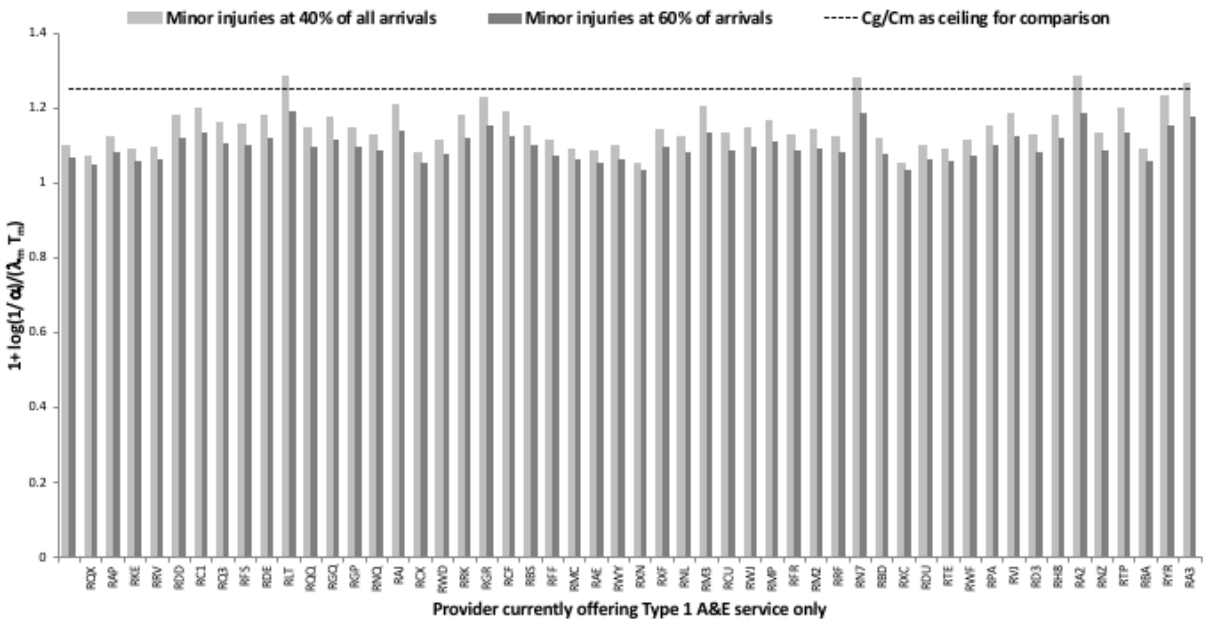

Figure 4 Most providers with Type-1-only A&Es would reduce cost by splitting off Type-3 A&Es. *Source: NHS.*

5.2. Mobile Type-3 Facilities for Minor Injuries including Weekend Drunkenness

A special case of minor injuries is "acute alcohol intoxication" where the A&E still has to examine the patient for any other symptoms or injuries. Given routine drunkenness in city centers on weekends, some cities are trying mobile A&E units dubbed 'booze buses' to respond quickly to drunken patients as well as to not let service deteriorate in the regular A&E. Such a unit, the Alcohol Recovery Centre, is a 65-foot truck trailer equipped with several beds, a waiting area and showers. If the percentage of patients on a Friday or Saturday night in a town center were, say, 60% of all patients for the nearest A&E as discussed above, the case for splitting off a mobile Type-3 A&Es for weekends would be strong for that town centre (Figure 4, dark grey bars). On nights near New Year, the number can go up to 70% according to Simon Stevens, Chief Executive of NHS England. An alternative is a holding area or "drunk tank", a static version of the booze bus.

Yet another example of a mobile facility is the ambulance itself becoming a Type-3 A&E so that the patient can be discharged without being brought into the A&E if the injuries are minor. The frontline staff would share photographs or hold video consultations with colleagues based in clinical

hubs in the control room; all 4,000 frontline staff already have iPads to access patient records ("Frontline staff given iPads to access patient records at scene", Evening Standard, 30 May, 2018). London Ambulance Service responded to 1.2 million incidents in 2017-18, and wants to reduce the proportion of people taken to the A&E from 63% to 53%.

5.3. Splitting off both Type-2 and Type-3 A&Es from Type-1 A&Es

Assume that a pooled Type-1 A&E has three distinct types of arrivals: general patients (with non-minor injuries) at arrival rate λ_g, specialty patients (with non-minor injuries) at arrival rate λ_s, and minor-injury patients at arrival rate λ_m. If we have to consider splitting off both a Type-2 and a Type-3 A&E from a Type-1-only A&E, there are four possible configurations: (i) staying as Type-1 only, (ii) fully split with Type-2 and Type-3 A&Es, (iii) with only Type-3 split off, and (iv) with only Type-2 split off. The optimal configuration is then characterized by:

PROPOSITION 3. *(a) If* $\left(\frac{c_g}{c_m}-1\right)\lambda_m < \frac{1}{\kappa T_m}\ln\left(\frac{1}{\alpha}\right)$ *and* $\frac{c_g}{c_s} < 1-p_s$*, then system (ii) has the least cost.*

(b) If $\left(\frac{c_g}{c_m}-1\right)\lambda_m \geq \frac{1}{\kappa T_m}\ln\left(\frac{1}{\alpha}\right)$ *and* $\frac{c_g}{c_s} \geq 1-p_s$*, then system (i) has the least cost.*

(c) If $\left(\frac{c_g}{c_m}-1\right)\lambda_m \geq \frac{1}{\kappa T_m}\ln\left(\frac{1}{\alpha}\right)$ *and* $\frac{c_g}{c_s} < 1-p_s$*, then system (iv) has the least cost.*

(d) If $\left(\frac{c_g}{c_m}-1\right)\lambda_m < \frac{1}{\kappa T_m}\ln\left(\frac{1}{\alpha}\right)$ *and* $\frac{c_g}{c_s} \geq 1-p_s$*, then system (iii) has the least cost.*

Proposition 3 shows that cost-effectiveness is determined by two comparisons: (1) whether $\left(\frac{c_g}{c_m}-1\right)\lambda_m$ is greater or smaller than $\frac{1}{\kappa T_m}\ln\left(\frac{1}{\alpha}\right)$ to indicate whether or not to retain minor injury patients with general patients, (2) whether $\frac{c_g}{c_s}$ is greater or smaller than $1-p_s$ to indicate whether or not to retain specialty patients with general patients. Note that $\left(\frac{c_g}{c_m}-1\right)\lambda_m < \frac{1}{\kappa T_m}\ln\left(\frac{1}{\alpha}\right)$ if and only if it is cost-effective to split off a Type-3 A&E, which contrasts configurations (ii) and (iii) above to configurations (i) and (iv). The condition depends on the parameters for Type-1 patients and Type-3 patients, but is independent of the parameters related to the Type-2 patients. Furthermore, $\frac{c_g}{c_s} < 1-p_s$ if and only if it is cost-effective to split off a Type-2 A&E, which contrasts configurations (ii) and (iv) to (i) and (iii). The comparison depends on the parameters related the Type-1 patients and Type-2 patients, but is independent of the parameters related to the Type-3 patients. Thus, we have:

COROLLARY 1. *The decision to split off a Type-3 A&E for minor injuries can be made independently of the decision to split off a Type-2 A&E for a particular specialty from the same Type-1 A&E.*

For this reason, we exclude Type-3 A&Es from consideration in the following section to focus only on specialty patients of a particular type, noting that a hospital can have multiple Type-2 facilities for different specialties.

6. Case 3: Reconfiguring an A&E Network: The case of two hospitals

We focus on a minimal network with only two hospitals and with the practical setting of $T_g \leq T_s$ to show that mergers are not always cost-effective. Even in this minimal setting, there are six possible network configurations based on (1) whether or not to close the A&Es in one hospital and, (2) whether a pooled system (Type 1 only) or a split system (Type 2 along with Type 1) is used. If the A&Es in two hospitals are not merged, then each hospital has the choice of having a pooled or split system with regard to Type-1 and Type-2 patients, giving us Systems (I), (II), (V), and (VI) in Table 2, and if the A&Es in two hospitals are merged, then a pooled or split system may be created subsequently, giving us Systems (III) and (IV) respectively in Table 2.

System	(I)	(II)	(III)	(IV)	(V)	(VI)
Hospital 1 A&E	P	S	-	-	P	S
Hospital 2 A&E	P	S	-	-	S	P
Merged A&E service	-	-	P	S	-	-

Table 2 Six possible network configurations for a two-hospital A&E network ("P" for pooled and "S" for split).

System (III) dominates System (I) by always having a lower cost because of the pooling effect in queueing and, likewise, System (IV) dominates System (II). Therefore, we drop Systems I and II from consideration. As we evaluate costs for alternative reconfigurations in this stylized model for policy insight, we note that for operational decisions other factors will have to be considered. For example, some patients may have to travel more compared if there is a merger, unless this is

in an urban area with the hospitals close to each other. Mergers also reduce choice for patients to a point that the Competition and Markets Authority would not allow it. Such location-specific factors notwithstanding, we consider only the cost of A&E service for systemwide comparisons.

6.1. Optimal network configurations

Recalling the notation already introduced – T_g, T_s, and α – we introduce superscript i ($i=1,2$) for hospital i in the network. Then, for hospital i the arrival rate for general (Type-1) patients is λ_g^i, the arrival rate for speciality (Type-2) patients is λ_s^i, and the admission rate for specialty patients is p_s^i. We will use some intermediate parameters for our analysis:

$$p_s = \frac{\lambda_s^1 p_s^1 + \lambda_s^2 p_s^2}{\lambda_s^1 + \lambda_s^2}; \tag{10}$$

$$\delta = \frac{c_g}{c_s} + p_s - 1 - \frac{1}{\lambda_s^1 + \lambda_s^2}\left(\frac{1}{\kappa T_g} - \frac{1}{\kappa T_s}\right)\ln\left(\frac{1}{\alpha}\right); \tag{11}$$

$$\beta^i = \frac{c_g}{c_s} + p_s^i - 1 + \frac{c_g}{c_s}\frac{1}{\kappa T_g \lambda_s^i}\ln\left(\frac{1}{\alpha}\right) + \frac{1}{\kappa T_s \lambda_s^i}\ln\left(\frac{1}{\alpha}\right),\ i=1,2. \tag{12}$$

$$\gamma^i = \frac{c_g}{c_s} + p_s^i - 1 - \left(1 + \frac{c_g}{c_s}\right)\frac{1}{\kappa T_g \lambda_s^i}\ln\left(\frac{1}{\alpha}\right),\ i=1,2. \tag{13}$$

These parameters are useful for characterizing the configuration landscape of the two-hospital A&E network. Of particular interest are the two parameters:

$$\delta^i = \frac{c_g}{c_s} + p_s^i - 1 - \frac{1}{\lambda_s^i}\left(\frac{1}{\kappa T_g} - \frac{1}{\kappa T_s}\right)\ln\left(\frac{1}{\alpha}\right),\ i=1,2. \tag{14}$$

Taking $T_g \le T_s$ as before, and that the status quo is that the A&Es of both hospitals operate separately, we have the following from Proposition 1: (1) If $\delta^1 \ge 0$ and $\delta^2 \ge 0$, then it is optimal to have a split system at each of the two hospitals; (2) if $\delta^1 \ge 0$ and $\delta^2 < 0$, then it is optimal to have a split system at Hospital 1 and a pooled system at Hospital 2; (3) If $\delta^1 < 0$ and $\delta^2 \ge 0$, then it is optimal to have a pooled system at Hospital 1 and a split system at Hospital 2; and (4) if $\delta^1 < 0$ and $\delta^2 < 0$, then it is optimal to have a pooled system at each of the two hospitals. The main results in this section are summarized in the proposition below and in Table 3:

PROPOSITION 4. *Assuming that $T_g \le T_s$ and one of the two A&Es can be considered for closure as a result of their merger:*

(A) If $\delta^1 \geq 0$ and $\delta^2 \geq 0$, then it is optimal to close the A&E in one hospital and operate a split system in the merged organization.

(B) If $\delta^1 \geq 0$ and $\delta^2 < 0$, then it is optimal to have

(i) A merged pooled system if and only if $\delta < 0$ and $\gamma^1 < 0$.

(ii) A merged-split system if and only if $\delta \geq 0$ and $\beta^2 \geq 0$.

(iii) A split system for Hospital 1 and a pooled system for Hospital 2 if and only if (1) $\delta < 0$ and $\gamma^1 \geq 0$ or (2) $\delta \geq 0$ and $\beta^2 < 0$.

(iv) A merged-split system if $p_s^1 = p_s^2$.

(C) If $\delta^1 < 0$ and $\delta^2 \geq 0$, then it is optimal to have

(i) A merged pooled system if and only if $\delta < 0$ and $\gamma^2 < 0$.

(ii) A merged split system if and only if $\delta \geq 0$ and $\beta^1 \geq 0$.

(iii) A pooled system for Hospital 1 and a split system for Hospital 2 if and only if either (1) $\delta < 0$ and $\gamma^2 \geq 0$ or (2) $\delta \geq 0$ and $\beta^1 < 0$

(iv) A merged split system if $p_s^1 = p_s^2$.

(D) If $\delta^1 < 0$ and $\delta^2 < 0$, then it is optimal to have

(i) A merged split system if $\delta \geq 0$.

(ii) A merged pooled system if $\delta < 0$.

The main insight drawn from Proposition 4 is that reconfiguration in line with the Keogh recommendations by merging the A&Es of several hospitals and then creating multiple Type-2 specialties is not necessarily cost-optimal. Indeed, the decision is rather nuanced as reflected by the main results: (1) Even if closing the A&E in one hospital is an option, it is not necessarily optimal to do so as pooling does not always reduce costs. (2) Even if it is optimal to merge the two A&E's into one hospital, the optimal configuration for the merged system could be split or pooled. (3) Whether or not the A&E in one hospital should be closed and whether a merged split or pooled system is used depends on the values of such parameters as the cost ratio $\frac{c_g}{c_s}$, the admission rate p_s^i, arrival rates λ_s^i, and waiting time targets T_g and T_s. (4) When $p_s^1 = p_s^2$, the two hospitals' A&Es

are always merged and the merged organization operates with a split system if $\delta \geq 0$ and a pooled system if $\delta < 0$. This shows that if the A&Es of the two hospitals are the same in terms of admission rates for Type-2 patients, then it is optimal to merge them. (5) When $\delta \geq 0$, a split system must be employed by the two hospitals' A&Es if they are merged. (6) When $\delta < 0$, a pooled system must be employed by the two hospitals' A&Es if they are merged. This can be viewed as an extension of Proposition 1 in an A&E network setting (Table 3).

	$\delta^2 \geq 0$	$\delta^2 < 0$
$\delta^1 \geq 0$	(S, S) $\longrightarrow$ [S]	(S, P) $\longrightarrow$ [P] iff $\delta < 0$ and $\gamma^1 < 0$ $\longrightarrow$ [S] iff $\delta \geq 0$ and $\beta^2 \geq 0$ $\longrightarrow$ (S, P) iff $\delta \geq 0$ and $\beta^2 < 0$ or $\delta < 0$ and $\gamma^1 \geq 0$
$\delta^1 < 0$	(P, S) $\longrightarrow$ [P] iff $\delta < 0$ and $\gamma^2 < 0$ $\longrightarrow$ [S] iff $\delta \geq 0$ and $\beta^1 \geq 0$ $\longrightarrow$ (P, S) iff $\delta \geq 0$ and $\beta^1 < 0$ or $\delta < 0$ and $\gamma^2 \geq 0$	(P, P) $\longrightarrow$ [S] iff $\delta \geq 0$ $\longrightarrow$ [P] iff $\delta < 0$

Table 3 Optimal A&E reconfigurations with two hospitals (P=pooled and S=split)

6.2. Sensitivity analysis of network configurations

Proposition 4 shows that the signs for the parameters δ, δ^1, δ^2, β^1, β^2, γ^1, and γ^2 play an important role in determining the optimal configuration for a given A&E network. In real life, λ_s^1 and λ_s^2 change dynamically by hour during the day, by day during the week, and by month during the year. Assuming that all other parameters are unchanged, we can explore how the optimal A&E network configuration changes when demand parameters λ_s^1 and λ_s^2 change; for example, the conditions under which the two hospitals' A&Es should be merged and how the optimal network configuration evolves with a change in the arrival rate for Type-2 patients.

PROPOSITION 5. *Assume that $T_g \leq T_s$, that the two hospitals' A&Es currently operate separately, and that the A&E of one hospital is considered for closure.*

(a) If $\frac{c_g}{c_s}+p_s^1-1$ *and* $\frac{c_g}{c_s}+p_s^2-1$ *have the same sign, then it is optimal to merge the two hospitals' A&Es.*

(b) If $\frac{c_g}{c_s}+p_s^i-1<0$*, then increasing* λ_s^i *reinforces the pooling strategy.*

(c) If $\frac{c_g}{c_s}+p_s^i-1\geq 0$*, then increasing* λ_s^i *reinforces the splitting strategy.*

This leads to our second important managerial insight: when $\frac{c_g}{c_s}+p_s^1-1$ and $\frac{c_g}{c_s}+p_s^2-1$ have the same sign, the optimal configuration for the two hospitals' A&Es is a merger, extending the fourth result in the previous subsection, where $p_s^1=p_s^2$. Depending on the population demographics in the two hospitals' catchment areas, say proportion of geriatric people, we note that if the two hospitals' A&Es have very similar admission rates to hospital for Type-2 patients, they should be merged. This is in line with the queuing literature that recommends pooling to reduce costs when customers are homogeneous (van Dijk 2008). Moreover, a merged split system as envisioned in the Keogh report becomes more attractive when either λ_s^1 or λ_s^2 is increasing. This reinforces the same observation for A&E departments in a single hospital.

When $\frac{c_g}{c_s}+p_s^1-1$ and $\frac{c_g}{c_s}+p_s^2-1$ have opposite signs, the two hospitals' A&Es may or may not be merged. On one hand, when $\frac{c_g}{c_s}+p_s^i-1\geq 0$, an increase in λ_s^i makes a split system more attractive for the A&E of hospital i. On the other hand, when $\frac{c_g}{c_s}+p_s^i-1<0$, an increase in λ_s^i makes a pooled system more attractive for hospital i. This observation cannot be made for the A&E in a single-hospital setting because when $\frac{c_g}{c_s}+p_s^i-1<0$, it is optimal to have a pooled system for the A&E in hospital i no matter how large λ_s^i is; i.e. a low admission rate p_s^i and low cost ratio of c_g/c_s outweigh a large arrival rate of λ_s^i.

The results in Proposition 5 are derived from a preliminary result in the proof in **Appendix B**, which shows that the $(\lambda_s^1,\lambda_s^2)$ is divided into different regions where the same network configuration is used. This result resembles the well-known *two-dimensional switching curve policy*, an extension of the one-dimensional threshold policy in the operations management literature (Porteus 2002).

We further investigate the sensitivity of the A&E network configuration by changing admission rates p_s^1 and p_s^2 but fixing all other parameters. Similar to Propositions 5, we obtain some complementary comparative static results on the optimal network configuration with a change in the

admission rates for specialty patients and the conditions related to the admission rates under which the A&E systems should be merged or remain separate.

PROPOSITION 6. *Assume that $T_g \leq T_s$, that the two hospitals' A&Es currently operate separately, and that the A&E in one hospital is considered for closure.*

(a) If p_s^i is increased, then it is more attractive to employ a splitting strategy.

(b) If both p_s^1 and p_s^2 are sufficiently close to 1, then it is optimal to have a merged split system.

(c) If both p_s^1 and p_s^2 are sufficiently close to 0, then it is optimal to have a merged pooled system.

(d) If one of p_s^1 and p_s^2 is sufficiently close to 1 and the other is sufficiently close to 0, then it is optimal to run the two A&Es separately.

As with Proposition 5, the results in Proposition 6 are derived from the so-called switching curve policies in the network configuration landscape in terms of admission rates. Furthermore, Proposition 6 implies interesting monotone properties, e.g., an increase in p_s^i makes a split system more attractive than a pooled system. This is an extension of the equivalent observation for a single-hospital A&E setting.

Overall, on one hand, a merged system is preferable when both p_s^i are relatively large (a merged split system) or both are relatively small (a merged pooled system). On the other hand, the two hospitals' A&Es operate separately when one has a relatively high admission rate for Type-2 patients and the other has a relatively low admission rate for Type-2 patients. Thus, the pooling effect increases with customer homogeneity and diminishes with customer heterogeneity (van Dijk 2008).

6.3. Illustrations with City of Leicester and Hammersmith, London

After a merger of two hospitals' A&Es in the city of Leicester, a dedicated A&E (Type 2) was created for geriatric patients (Conroy et al. (2014)). We wish to verify whether the merger and creation of a dedicated Type-2 A&E is supported by our analysis. Before the merger, the aggregated arrival rate in the two units for patients aged 65 and older was 6 per hour and the aggregated admission rate was 36%. Recall that $T_g = 4$, $T_s = 12$, and $\alpha = 5\%$. For further analysis, we take

the ratio c_g/c_s as a parameter. Regarding the arrival and admission rates for geriatric patients at the two A&Es, we assume that $\lambda_s^1 + \lambda_s^2 = 6$ and $(p_s^1\lambda_s^1 + p_s^2\lambda_s^2)/(\lambda_s^1 + \lambda_s^2) = 36\%$, which implies that $\lambda_s^2 = 6 - \lambda_s^1$ and $p_s^2 = (0.36(\lambda_s^1 + \lambda_s^2) - p_s^1\lambda_s^1)/\lambda_s^2$. Thus, we vary the values for λ_s^1 and p_s^1 and obtain values for λ_s^2 and p_s^2 using these relationships. Based on Proposition 4, we obtain the results on the optimal A&E configurations prior to and following the merger shown in Table 4.

In all scenarios, our analysis supports the merger. However, whether or not to establish a separate geriatric A&E depends on the relative cost c_g/c_s. When geriatric specialists are much more expensive than generalists (e.g. $c_g/c_s = 0.6$), a pooled system is preferable after the merger; otherwise, a split system is preferable (Table 4).

Consider another example of five A&Es in London in the Hammersmith area. It is natural to consider possible mergers of A&Es. Five hospitals, with codes RQM, CXC, SMH, RJ1, and RRV, are located in close proximity to each other. Based on Proposition 4, we analyze whether it would be beneficial to merge any two of these five A&Es in terms of reducing costs and increasing service quality.

Data for the five A&Es is presented in Table 5. Following Proposition 1 and using the data in Table 5, we can see that the optimal configuration for all five A&Es is a split system provided that the admission rate for geriatric patients is equal to or greater than 20%. The NHS data shows that the national average admission rate for all A&E patients is around 20% and the admission rate for geriatric patients is often significantly higher than this average.

Following Proposition 4, we can test whether any pair of the five A&Es should be merged and if merged, whether Type-2 A&Es (a split system) should be created. For each A&E, we use three different values for the admission rate of geriatric patients: 0.35, 0.50 and 0.65. Thus, for each pair of A&Es we generate nine scenarios. The results are shown in Table 6 for $c_g/c_s = 0.7$ and $c_g/c_s = 0.6$, where the admission rates for two A&Es are shown in the first row, 'S' indicates a merged and split system, and 'P' indicates a merged and pooled system. Due to the high admission rates in all five A&Es, our analysis shows that in all scenarios, a merger would be beneficial and that a split system would be preferable to a pooled system when c_g/c_s and/or the admission rate for geriatric patients is higher.

$(\lambda_s^1, \lambda_s^2)$	(p_s^1, p_s^2)	$c_g/c_s = 0.6$	$c_g/c_s = 0.7$	$c_g/c_s = 0.8$	$c_g/c_s = 0.9$
(4.5, 1.5)	(0.30,0.54)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(P, S) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]
	(0.33,0.45)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]
	(0.36,0.36)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [S]	(S, P) $\longrightarrow$ [S]
	(0.39,0.27)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [S]	(S, P) $\longrightarrow$ [S]
	(0.42,0.18)	(P, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [S]	(S, P) $\longrightarrow$ [S]
(4.0, 2.0)	(0.30,0.48)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(P, S) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]
	(0.33,0.42)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]
	(0.36,0.36)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]
	(0.39,0.30)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [S]	(S, P) $\longrightarrow$ [S]
	(0.42,0.24)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [S]	(S, P) $\longrightarrow$ [S]
(3.5, 2.5)	(0.30,0.44)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(P, S) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]
	(0.33,0.40)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(P, S) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]
	(0.36,0.36)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]
	(0.39,0.32)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]
	(0.42,0.28)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [S]	(S, P) $\longrightarrow$ [S]
(3.0, 3.0)	(0.30,0.42)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(P, S) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]
	(0.33,0.39)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(P, S) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]
	(0.36,0.36)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]
	(0.39,0.33)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]
	(0.42,0.30)	(P, P) $\longrightarrow$ [P]	(P, P) $\longrightarrow$ [P]	(S, P) $\longrightarrow$ [S]	(S, S) $\longrightarrow$ [S]

Table 4 **Results for the Leicester merger and A&E configurations.**

Code	RQM	CXC	SMH	RJ1	RRV
λ_g	30.81	5.81	11.63	23.05	21.37
λ_s	2.87	0.80	1.60	2.09	2.23
p_s	50%	50%	50%	50%	50%

Table 5 **A&E data for five London hospitals (RQM, CXC, SMH, RJ1, and RRV).**

	0.35,0.35	0.35,0.50	0.35,0.65	0.50,0.35	0.50,0.50	0.50,0.65	0.65,0.35	0.65,0.50	0.65,0.65
RQM,CXC	P	P	P	S/P	S/P	S/P	S	S	S
RQM,SMH	P	P	S/P	S/P	S/P	S	S	S	S
RQM,RJ1	P	S/P	S/P	S/P	S/P	S	S	S	S
RQM,RRV	P	S/P	S/P	S/P	S	S	S	S	S
CXC,SMH	P	P	S/P	P	P	S/P	P	S/P	S
CXC,RJ1	P	P	S/P	P	S/P	S	P	S/P	S
CXC,RRV	P	P	S	P	S/P	S	P	S/P	S
SMH,RJ1	P	P	S/P	P	S/P	S	S/P	S	S
SMH,RRV	P	S/P	S/P	P	S/P	S	S/P	S	S
RJ1,RRV	P	S/P	S/P	S/P	S/P	S	S/P	S	S

Table 6 Alternative configurations with $c_g/c_s = 0.7$ and $c_g/c_s = 0.6$: a single letter implies the same configuration (Split or Pooled) at both costs levels, two letters, S/P or P/S, imply different configurations.

7. Conclusion

We used stylized modeling to take a closer look at the reconfiguration of the A&E system in NHS England as proposed by the Keogh report of 2013. *First*, we investigated when a Type-1 A&E should consider splitting off a Type-2 A&E. The results show how increasing volumes of specialty patients tilt the balance in favor of Type-2 A&Es, thus giving support to the idea of 'mega-centres' of multiple Type-2 A&Es in densely-populated urban areas. One implication was the case for splitting off geriatric A&Es. *Second*, we repeated this analysis for splitting off Type-3 services. These results support the Keogh recommendation of having many Type-3 A&Es for minor injuries in urban centres. We also considered temporary facilities for weekend drunkenness. *Finally*, we considered the optimal configuration of two hospitals' A&E facilities and showed there were cases where a merger-induced pooling (and then splitting off multiple Type-2 services in line with the Keogh report) would not be optimal. We illustrated this with two hospitals in Leicester that merged in 2011 and with five London hospitals in close proximity to each other.

Our work is aimed at healthcare policy for reconfiguring A&E facilities systemwide as envisioned

by the Keogh report. Modeling has been successful at the hospital level (Green and Kolesar 2004, Musafee 2016a and Musafee 2016b), now the focus needs to shift to the system-wide level to aid healthcare policy given the rising costs of healthcare in general, not just of A&E services, in most countries.

Thus, one extension of the work in this paper is to use more detailed modeling, for instance, to further evaluate candidate reconfiguration identified by stylized models. Situation-specific models are needed to additionally address increased travel times and reduced choice for patients. At the hospital level specifically, the literature already looks at how to arrange patient flow in an A&E using 'triage' to assign patients to different groups – including prioritizing as well as where to send the patient. Williams (2006) argues that a fast-track lane for low-acuity patients reduces overcrowding given that three quarters of A&E patients are non-urgent. Flinders Medical Center in Australia has implemented a new method whereby patients are streamed based on their likelihood of being admitted to hospital, resulting in a significant reduction in average waiting times (King et al. 2006).

Heathcare policy would benefit also from empirical work within and across hospitals to obtain the parameters that stylized and detailed models would require. Details of case studies, such as those of the Leicester A&E unit (Conroy et al. 2014), would also be helpful in identifying the benefits and challenges raised by mergers.

Much work remains, however, as regards system-wide modeling of A&E services. The paper by Xu and Chan (2016) on proactive policies for preventing buildup of excessive waiting times by diverting patients to other A&E facilities is a step in this direction. See also Henderson (2008) for an overview of the challenges and the use of Approximate Dynamic Programming in this regard. As was the case with aggregate planning with supply chain models, a mix of mathematical programming and queuing models could be valuable for matching (forecasted) demand for A&E services to propose an optimal A&E network. We hope this paper has provided a start.

Acknowledgement. *We are grateful to the Senior Editor and the two reviewers for providing constructive and useful comments which have significantly improved the presentation of this paper.*

References

Abate, J., G.L. Choudury, W. Whitt. 1995. Exponential approximations for tail probabilities in queues I: waiting times. *Operations Research*, 43(5): 885–901.

Abate, J., G.L. Choudhury, W. Whitt. 1996. Exponential approximations for tail probabilities in queues II: Sojourn time and workload. *Operations Research* 44 (5) 758–763.

Age UK. 2017. Later Life in the United Kingdom. Available at `http://www.ageuk.org.uk/Documents/EN-GB/Factsheets/Later_Life_UK_factsheet.pdf?dtrk=true` (accesssed on February 20, 2019).

Allon, G., A. Federgruen. 2008. Service competition with general queueing facilities. *Operations Research* 56(4): 827–849.

Andradottir, S., H. Ayhan, D.G. Down. 2017. Resource pooling in the presence of failures: Eficiency versus risk. *European Journal of Operational Research* 256(1): 230–241.

Armony, M., S. Israelit, A. Mandelbaum, Y.N. Marmor, Y. Tseytlin, G.B. Yom-Tov. 2015. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* 5(1): 146–194.

Benjaafar, S. 1995. Performance bounds for the effectiveness of pooling in multi-processing systems. *European Journal of Operational Research* 87(2): 375–388.

Burke, P.J. 1956. The output of a queuing system. *Operations Research* 4(6): 699–704.

Cachon G., C. Terwiesch. 2009. *Matching Supply with Demand: An Introduction to Operations Management.* McGraw-Hill Education, Singapore.

Cawston, T., A. Haldenby, N. Seddon. 2012. Healthy competition, White Paper.

Chan, C.W., V.F. Farias, N. Bambos, G.J. Escobar. 2012. Optimizing intensive care unit discharge decisions with patient readmissions. *Operations research* 60(6): 1323–1341.

Chan, C.W., G. Yom-Tov, G.J. Escobar. 2014. When to use speedup: An examination of service systems with returns. *Operations Research* 62(2): 462–482.

Conroy, S.P., K. Ansari, M. Williams, E. Laithwaite, B. Teasdale, J. Dawson, S. Mason, J. and Barnerjee. 2014. A controlled evaluation of comprehensive geriatric assessment in the emergency department: the 'Emergency Frailty Unit'. *Age and Ageing* 43: 109–114.

Cooke, M.W., S. Wilson, S. Pearson. The effect of a separate stream for minor injuries on accident and emergency department waiting times. *Emergency Medicine Journal* 19(1): 28–30.

Department of Health. 2016. *NHS Stnadard Contract 2016/17 Particular (Full Length)* `https://www.england.nhs.uk/wp-content/uploads/2016/02/2-full-lngth-16-17-partics.pdf` (accessed September 4, 2017).

Geddes, L. 2013. Solving an age-old problem. *New Scientist* (17 August) 219(2930): 8–9.

Green, L.V. 2012. OM Forum – The vital role of operations analysis in improving healthcare delivery. *Manufacturing & Service Operations Management* 14(4): 488–494.

Green, L.V., P.J. Kolesar. 2004. Improving emergency responsiveness with management science. *Management Science* 50(8): 1001–1014.

Green, L.V., S. Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Operations Research* 56(6): 1526–1538.

Henderson, S.G. 2011. Operations research tools for addressing current challenges in emergency medical services. *Wiley Encyclopedia of Operations Research and Management Science*, Chichester, UK.

Ieraci, S., E. Digiusto, P. Sonntag, L. Dann, D. Fox. 2008. Streaming by case complexity: Evaluation of a model for emergency department Fast Track. *Emergency Medicine Australas* 20: 241–249.

Jiang, H., Z. Pang, S. Savin. 2012. Performance-Based Contracts for Outpatient Medical Services, *Manufacturing & Service Operations Management* 14(2): 654–669.

King, D.L., D.I. Ben-Tovim, J. Bassham. 2006. Redesigning emergency department patient flows: Application of lean thinking to health care. *Emergency Medicine Australasia* 18: 391–397.

Mandelbaum, A., M.I. Reiman. 1998. On pooling in queueing networks. *Management Science* 44(7): 971-981.

Mayhew, L., D. Smith. 2008. Using queuing theory to analyse the Governments 4-h completion time target in Accident and Emergency departments. *Health Care Management Science* 11: 11–21.

Mirasol, N.M. 1963. Letter to the Editor–The output of an $M/G/\infty$ queuing system is Poisson. *Operations Research*, 11(2), 282–284.

Mustafee, N. 2016. *Operational research for emergency planning in healthcare: Volume 1.* Palgrave Macmillan, New York.

Musafee, N. 2016. *Operational research for emergency planning in healthcare: Volume 2*. Palgrave Macmillan, New York.

NHS England. 2013. *Transforming urgent and emergency care services in England, End-of-Phase 1 report*, Urgent and emergency care review team, Leeds, UK. Available at `http://www.nhs.uk/NHSEngland/keogh-review/Documents/UECR.Ph1Report.FV.pdf` (accesssed on February 20, 2019).

NHS England. 2017. Winter Daily Situation Reports. Available at `https://www.england.nhs.uk/statistics/statistical-work-areas/winter-daily-sitreps/` (accesssed on February 20, 2019).

NHS England. 2017. A&E Attendances and Emergency Admissions. Available at `https://www.england.nhs.uk/statistics/statistical-work-areas/ae-waiting-times-and-activity/` (accesssed on February 20, 2019).

NHS England. 2018. A&E attendances and emergency admissions. Available at `https://www.england.nhs.uk/statistics/statistical-work-areas/ae-waiting-times-and-activity/ae-attendances-and-emergency-admissions-2017-18/` (accesssed on February 20, 2019).

Porteus, E.L. 2002. *Foundation of Stochastic Inventory Theory*. Stanford University Press, Stanford.

Propper, C., M. Sutton, C. Whitnall, F. Windmeijer. 2008. Did "targets and terror" reduce waiting times in England for hospital care? *The B.E. Journal of Economic Analysis and Policy* 8(2): 1935–1682.

Saghafian, S., G. Austin, S.J. Traub. 2015. Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering* 5(2): 101–123.

Saghafian, S., W.J. Hopp, M.P. Van Oyen, J.S. Desmond, S.L. Kronick. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* 60(5): 1080–1097.

Saghafian, S., W.J. Hopp, M.P. Van Oyen, J.S. Desmond, S.L. Kronick. 2014. Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacuring & Service Operations Management* 16(3): 329–345.

Smith, D., W. Whitt. 1981. Resource sharing for efficiency in traffic Systems. *Bell System Technical Journal* 60(13): 39–55.

Song, H., A.L. Tucker, K.L. Murrell. 2015. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* 61(12): 3032–3053.

van Dijk, N.M., E. van der Sluis. 2008. To pool or not to pool in call centres. *Production and Operations Management* 17 (3): 296–305.

Williams, M. 2006. *Hospitals and Clinical Facilities, Processes and Design for Patient Flow, in Patient Flow: Reducing Delay in Healthcare Delivery*. Springer, New York.

Wright, P., G. Tang, S. Ilief, D. Lee. 2013. The impact of a new emergency admission avoidance system for older people on length of stay and same-day discharges. *Age and Ageing* 0: 1–6.

Xu, K., C.W. Chan. 2016. Using future information to reduce waiting times in the emergency department via diversion, *Manufacuring & Service Operations Management* 18(3): 314–331.

Whitt, W. 1993. Approximations for the GI/G/m queue. Production and Operations Management, 2(2): 114–161.

Whitt, W. 1999. Partitioning customers into service groups. *Management Science* 45 1579–1592.

Appendix A: Justification of approximation of aggregate service capacity requirement

We follow the various approximations for the $M/G/s$ queue in Whitt (1999), where heavy traffic is assumed. Let λ be the arrival rate, m_i be the i-th moment for the service time, $\tau_s^2 = \frac{m_2 - m_1^2}{m_1^2}$ the square coefficient of variation, $\nu = \frac{1}{m_1}$ the mean service rate per server, $\rho = \frac{\lambda}{s\nu}$ the traffic intensity, and W_q the waiting time in the queue. Whitt (1999) shows that

$$E[W_q(M/M/s)|W_q(M/M/s) > 0] = \frac{1}{s\nu(1-\rho)} \tag{15}$$

$$E[W_q(M/G/s)|W_q(M/G/s) > 0] \approx \frac{1+\tau_s^2}{2} E(W_q[M/M/s)|W_q(M/M/s) > 0] \tag{16}$$

$$= \frac{(1+\tau_s^2)}{2s\nu(1-\rho)} \tag{17}$$

$$P(W_q(M/G/s) > 0) \approx P(W_q(M/M/s) > 0) \tag{18}$$

$$\approx \rho^{\sqrt{2(s+1)}-1}. \tag{19}$$

Thus

$$E[W_q(M/G/s)] = E[W_q(M/G/s)|W_q(M/G/s) > 0]P(W_q(M/G/s) > 0) \tag{20}$$

$$\approx \frac{1+\tau_s^2}{2} E[W_q(M/M/s)|W_q(M/M/s) > 0]P(W_q(M/G/s) > 0) \tag{21}$$

$$\approx \frac{1+\tau_s^2}{2} E[W_q(M/M/s)|W_q(M/M/s) > 0]P(W_q(M/M/s) > 0) \tag{22}$$

$$\approx \frac{(1+\tau_s^2)}{2s\nu(1-\rho)} \rho^{\sqrt{2(s+1)}-1}. \tag{23}$$

Whitt (1999) also shows that

$$P(W_q(M/G/s) > T) \approx P(W_q(M/G/s) > 0) e^{-\frac{T}{E[W_q(M/G/s)|W_q(M/G/s)>0]}} \tag{24}$$

$$\approx P(W_q(M/M/s) > 0) e^{-\frac{T}{E[W_q(M/G/s)|W_q(M/G/s)>0]}} \tag{25}$$

$$\approx \rho^{\sqrt{2(s+1)}-1} e^{-\frac{T}{E[W_q(M/G/s)|W_q(M/G/s)>0)]}} \tag{26}$$

$$\approx \rho^{\sqrt{2(s+1)}-1} e^{-\frac{2s\nu(1-\rho)}{(1+\tau_s^2)}T}. \tag{27}$$

We next follow the various approximations for the $M/G/s$ queue in Abate et al. (1996). Let V, W_q and W be the service time, waiting time and sojourn time random variables for the $G/GI/1$ queue. Theorem 1 in Abate et al. (1996) states the following (before Theorem 1, the authors also remark that the result can be extended to $G/GI/s$ easily). If $e^{\eta T} P(W_q > T) \to \alpha_1$ as $T \to \infty$, then $E[e^{\eta V}] < \infty$ and

$$e^{\eta T} P(W > T) \to \alpha_2 = \alpha_1 E\left[e^{\eta V}\right] > \alpha_1, \text{ as } T \to \infty.$$

As such, we have

$$P(W>T)\approx\alpha_1 e^{-\eta T}\approx\rho^{\sqrt{2(s+1)}-1}e^{-\frac{2s\nu(1-\rho)}{(1+\tau_s^2)}T}E\left[e^{\frac{2s\nu(1-\rho)}{(1+\tau_s^2)}V}\right]. \tag{28}$$

Note that when x is close to zero, $e^x\approx 1+x$. Therefore, when ρ is close to 1 (under the heavy traffic assumption), we have

$$\begin{aligned}
E\left[e^{\frac{2s\nu(1-\rho)}{(1+\tau_s^2)}V}\right] &\approx E\left[1+\frac{2s\nu(1-\rho)}{(1+\tau_s^2)}V\right] &(29)\\
&= 1+\frac{2s\nu(1-\rho)}{(1+\tau_s^2)}E[V] &(30)\\
&= 1+\frac{2s\nu(1-\rho)}{(1+\tau_s^2)}\frac{1}{\nu} &(31)\\
&\approx e^{\frac{2s(1-\rho)}{(1+\tau_s^2)}}. &(32)
\end{aligned}$$

This shows that

$$\begin{aligned}
P(W>T) &\approx \rho^{\sqrt{2(s+1)}-1}e^{-\frac{2s\nu(1-\rho)}{(1+\tau_s^2)}T}e^{\frac{2s(1-\rho)}{(1+\tau_s^2)}} &(33)\\
&= Ae^{-B(s\nu-\lambda)T}, &(34)
\end{aligned}$$

where

$$A=\rho^{\sqrt{2(s+1)}-1}e^{\frac{2s(1-\rho)}{(1+\tau_s^2)}},\quad B=\frac{2}{1+\tau_s^2}. \tag{35}$$

Here ρ is close to 1, which implies that $A\approx 1$.

Appendix B: Proofs

Proof of Proposition 2

Following an argument similar to that for Case 1, the minimum total cost for the dedicated system is

$$c_g\left(\lambda_g+\frac{1}{\kappa T_g}\ln\left(\frac{1}{\alpha}\right)\right)+c_m\left(\lambda_m+\frac{1}{\kappa T_m}\ln\left(\frac{1}{\alpha}\right)\right), \tag{36}$$

and the minimum total cost for the pooled system is

$$c_g\left(\lambda_g+\lambda_m+\frac{1}{\kappa T_g}\ln\left(\frac{1}{\alpha}\right)\right). \tag{37}$$

The difference in costs in the pooled and dedicated systems is $c_g\lambda_m-c_m\left(\lambda_m+\frac{1}{\kappa T_m}\ln\left(\frac{1}{\alpha}\right)\right)$. So, the dedicated system is more cost-effective than the pooled system if and only if $c_g\lambda_m\geq c_m\left(\lambda_m+\frac{1}{\kappa T_m}\ln\left(\frac{1}{\alpha}\right)\right)$. □

Proof of Proposition 3

The minimum total cost $C(.)$ for the systems (i), (ii), (iii), and (iv) respectively is:

$$C(i) = c_g \left[\lambda_g + \lambda_s + \lambda_m + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right] + c_s \left[p_s\lambda_s + \frac{1}{\kappa T_s} \ln\left(\frac{1}{\alpha}\right)\right]. \tag{38}$$

$$C(ii) = c_g \left[\lambda_g + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right] + c_s \left[\lambda_s + \frac{1}{\kappa T_s} \ln\left(\frac{1}{\alpha}\right)\right] + c_m \left[\lambda_m + \frac{1}{\kappa T_m} \ln\left(\frac{1}{\alpha}\right)\right]. \tag{39}$$

$$C(iii) = c_g \left[\lambda_g + \lambda_s + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right] + c_s \left[p_s\lambda_s + \frac{1}{\kappa T_s} \ln\left(\frac{1}{\alpha}\right)\right] + c_m \left[\lambda_m + \frac{1}{\kappa T_m} \ln\left(\frac{1}{\alpha}\right)\right]. \tag{40}$$

$$C(iv) = c_g \left[\lambda_g + \lambda_m + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right] + c_s \left[\lambda_s + \frac{1}{\kappa T_s} \ln\left(\frac{1}{\alpha}\right)\right]. \tag{41}$$

Define differences between the minimum total costs for any pair of systems (i), (ii), (iii), and (iv), as $D(k,m) = C(k) - C(m)$, where $k, m = i, ii, iii, iv$ and $k \neq m$ to obtain:

$$D(i,ii) = c_m \left[\left(\frac{c_g}{c_m} - 1\right)\lambda_m - \frac{1}{\kappa T_m} \ln\left(\frac{1}{\alpha}\right)\right] + c_s \left[\frac{c_g}{c_s} - (1-p_s)\right]\lambda_s, \tag{42}$$

$$D(i,iii) = c_m \left[\left(\frac{c_g}{c_m} - 1\right)\lambda_m - \frac{1}{\kappa T_m} \ln\left(\frac{1}{\alpha}\right)\right], \tag{43}$$

$$D(i,iv) = c_s \left[\frac{c_g}{c_s} - (1-p_s)\right]\lambda_s, \tag{44}$$

$$D(ii,iii) = -c_s \left[\frac{c_g}{c_s} - (1-p_s)\right]\lambda_s, \tag{45}$$

$$D(ii,iv) = -c_m \left[\left(\frac{c_g}{c_m} - 1\right)\lambda_m - \frac{1}{\kappa T_m} \ln\left(\frac{1}{\alpha}\right)\right], \tag{46}$$

$$D(iii,iv) = -c_m \left[\left(\frac{c_g}{c_m} - 1\right)\lambda_m - \frac{1}{\kappa T_m} \ln\left(\frac{1}{\alpha}\right)\right] + c_s \left[\frac{c_g}{c_s} - (1-p_s)\right]\lambda_s. \tag{47}$$

We obtain the two equalities

$$D(i,iii) = -D(ii,iv), \qquad D(i,iv) = -D(ii,iii). \tag{48}$$

to compare the cost-effectiveness of the four system configurations. The intuition behind these is as follows: The key difference between (i) and (iii) is whether or not to split off a Type-3 A&E, which is the same as that between (ii) and (iv). Similarly, the key difference between (i) and (iv) is whether or not to split off a Type-2 A&E, which is the same as that between (ii) and (iii). □

Proof of Proposition 4

The results in (a)–(d) follow from Proposition 1. Next, we prove the results in (e)–(h). Recall that as regards the total system cost, System (III) dominates System (I) and System (IV) dominates System (II). Therefore, we only need to compare Systems (III), (IV), (V), and (VI). The system costs are:

$$C(III) = c_g \left[\lambda_g^1 + \lambda_s^1 + \lambda_g^2 + \lambda_s^2 + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right] + c_s \left[p_s^1\lambda_s^1 + p_s^2\lambda_s^2 + \frac{1}{\kappa T_s} \ln\left(\frac{1}{\alpha}\right)\right] \tag{49}$$

$$C(IV) = c_g \left[\lambda_g^1 + \lambda_g^2 + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right] + c_s \left[\lambda_s^1 + \lambda_s^2 + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right] \tag{50}$$

$$C(V) = c_g \left[\lambda_g^1 + \lambda_s^1 + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right] + c_s \left[p_s^1 \lambda_s^1 + \frac{1}{\kappa T_s} \ln\left(\frac{1}{\alpha}\right)\right] \tag{51}$$

$$+c_g \left[\lambda_g^2 + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right] + c_s \left[\lambda_s^2 + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right] \tag{52}$$

$$C(VI) = c_g \left[\lambda_g^1 + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right] + c_s \left[\lambda_s^1 + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right] \tag{53}$$

$$+c_g \left[\lambda_g^2 + \lambda_s^2 + \frac{1}{\kappa T_g} \ln\left(\frac{1}{\alpha}\right)\right] + c_s \left[p_s^2 \lambda_s^2 + \frac{1}{\kappa T_s} \ln\left(\frac{1}{\alpha}\right)\right]. \tag{54}$$

We calculate the difference in costs between the two different systems. Again, we define $D(k,m) = C(k) - C(m)$, where $k, m = III, IV, V, VI$ and $k \neq m$, giving us:

$$D(III, IV) = c_s \left(\lambda_s^1 + \lambda_s^2\right) \delta, \tag{55}$$

$$D(IV, VI) = -c_s \lambda_s^2 \beta^2, \tag{56}$$

$$D(III, VI) = c_s \lambda_s^1 \gamma^1, \tag{57}$$

$$D(III, V) = c_s \lambda_s^2 \gamma^2, \tag{58}$$

$$D(IV, V) = -c_s \lambda_s^1 \beta^1. \tag{59}$$

(e) The result in (a) and Proposition 1 show that System (II) dominates Systems (I), (V), and (VI). We only need to compare Systems (II), (III), and (IV). Because System (II) is always dominated by System (IV), we only need to compare Systems (III) and (IV). Note that $D(III, IV) = c_s(\lambda_s^1 + \lambda_s^2)\delta \geq 0$ because $\delta^1 \geq 0$, $\delta^2 \geq 0$, $T_g \leq T_s$ and $\delta \geq \min\{\delta^1, \delta^2\}$. This implies that System (IV) is preferable to System (III). Hence, we have proved that it is optimal to close one A&E and have a merged split system (System (IV)).

(f) The result in (b) and Proposition 1 show that System (VI) dominates Systems (I), (II), and (V). We only need to compare Systems (III), (IV), and (VI). Looking at the conditions in (i), (ii), and (iii) and the values for $D(III, IV)$, $D(III, VI)$, and $D(IV, VI)$, it is easy to derive all the results in (i), (ii) and (iii). When $p_s^1 = p_s^2$, we have $\frac{c_g}{c_s} + p_s - 1 > 0$, $\delta \geq 0$ and $\beta^2 \geq 0$ because $\delta^1 \geq 0$ and $\delta \geq \delta^1$. Therefore, we have $D(III, IV) \geq 0$ and $D(IV, VI) \leq 0$, which shows that System (IV) is preferable.

(g) The result in (c) and Proposition 1 show that System (V) dominates Systems (I), (II), and (VI). We only need to compare Systems (III), (IV), and (V). Looking at the conditions in (i), (ii), and (iii) and the values for $D(III, IV)$, $D(IV, V)$, and $D(III, V)$, it is easy to derive all the results in (i), (ii) and (iii). When $p_s^1 = p_s^2$, we have $\frac{c_g}{c_s} + p_s - 1 > 0$, $\delta \geq 0$ and $\beta^1 \geq 0$ because $\delta^2 \geq 0$ and $\delta \geq \delta^2$. Therefore, we have $D(III, IV) \geq 0$ and $D(IV, V) \leq 0$, which shows that System (IV) is preferable.

(h) The result in (d) and Proposition 1 show that System (I) dominates Systems (II), (V), and (VI). We only need to compare Systems (I), (III), and (IV). However System (III) always dominates System (I), which indicates that we only need to compare System (III) and System (IV).

$$D(III, IV) = c_s \left(\lambda_s^1 + \lambda_s^2\right) \delta. \tag{60}$$

Therefore, if one of the two A&Es can be considered for closure, then it is optimal to close one A&E and operate a merged split system (System (IV)) if $\delta \geq 0$, and a merged pooled system (System (III)) if $\delta < 0$. This completes the proof. □

Proof of Proposition 5

(a) Based on Table 3, there are four cases where the optimal configuration for the two A&Es is to have one split system and another pooled system:

(1) $\delta^1 < 0$, $\delta^2 \geq 0$, $\delta \geq 0$, $\beta^1 < 0$,

(2) $\delta^1 < 0$, $\delta^2 \geq 0$, $\delta < 0$, $\gamma^2 \geq 0$,

(3) $\delta^1 \geq 0$, $\delta^2 < 0$, $\delta \geq 0$, $\beta^2 < 0$,

(4) $\delta^1 \geq 0$, $\delta^2 < 0$, $\delta < 0$, $\gamma^1 < 0$.

We aim to prove that none of the above four cases occurs. Note that for all four cases we must have $\frac{c_g}{c_s} + p_s^1 - 1 > 0$ and $\frac{c_g}{c_s} + p_s^2 - 1 > 0$ because $\frac{c_g}{c_s} + p_s^1 - 1$ and $\frac{c_g}{c_s} + p_s^2 - 1$ have the same sign and δ^1 and δ^2 have different signs. This shows that $\beta^1 > 0$ and $\beta^2 > 0$. Furthermore,

$$\delta = \frac{1}{\lambda_s^1 + \lambda_s^2} \left[\left(\frac{c_g}{c_s} + p_s^1 - 1 \right) \lambda_s^1 + \left(\frac{c_g}{c_s} + p_s^2 - 1 \right) \lambda_s^2 - \left(\frac{1}{\kappa T_g} - \frac{1}{\kappa T_s} \right) \ln \left(\frac{1}{\alpha} \right) \right] > 0$$

if one of δ^1 and δ^2 is non-negative. This shows that none of the above four cases occurs.

(b) We prove the result by looking at all cases in Table 3. Because of the symmetry property, we only need to prove the result for either p_s^1 or p_s^2.

If λ_s^1 is increased, then δ^1 and γ^1 are increased and β^1 is decreased but δ^2, β^2, and γ^2 remain unchanged. Note that $\delta^1 < 0$ and $\delta^1 < 0$ still hold even when λ_s^1 is increased. Thus, we only need to look at two cases: (1) $\delta^1 < 0$ and $\delta^2 \geq 0$ and (2) $\delta^1 < 0$ and $\delta^2 < 0$.

(1) Assume $\delta^1 < 0$ and $\delta^2 \geq 0$. If $\delta < 0$ and $\gamma^2 < 0$, then the optimal configuration for the A&E network is a merged pooled system. If we increase λ_s^1, then we have $\delta < 0$ and $\gamma^2 < 0$ because

$$\delta = \frac{1}{\lambda_s^1 + \lambda_s^2} \left[\left(\frac{c_g}{c_s} + p_s^1 - 1 \right) \lambda_s^1 + \left(\frac{c_g}{c_s} + p_s^2 - 1 \right) \lambda_s^2 - \left(\frac{1}{\kappa T_g} - \frac{1}{\kappa T_s} \right) \ln \left(\frac{1}{\alpha} \right) \right].$$

Thus, the optimal configuration for the A&E network remains a merged pooled system.

If $\delta < 0$ and $\gamma^2 \geq 0$, then the optimal configuration for the A&E network is to have a pooled system for A&E 1 and a split system for A&E 2. If we increase λ_s^1, then we have $\delta < 0$ and $\gamma^2 \geq 0$ because

$$\delta = \frac{1}{\lambda_s^1 + \lambda_s^2} \left[\left(\frac{c_g}{c_s} + p_s^1 - 1 \right) \lambda_s^1 + \left(\frac{c_g}{c_s} + p_s^2 - 1 \right) \lambda_s^2 - \left(\frac{1}{\kappa T_g} - \frac{1}{\kappa T_s} \right) \ln \left(\frac{1}{\alpha} \right) \right].$$

Thus, the optimal configuration for the A&E network remains a pooled system for A&E 1 and a split system for A&E 2.

If $\delta \geq 0$ and $\beta^1 \geq 0$, then the optimal configuration for the A&E network is to have a merged split system that does not contain the pooling strategy at all. Thus, it is obvious that a new optimal configuration with an increase in λ_s^1 would make the pooling strategy more attractive.

If $\delta \geq 0$ and $\beta^1 < 0$, then the optimal configuration for the A&E network is to have a pooled system for A&E 1 and a split system for A&E 2. If we increase λ_s^1, then β^1 is decreased and remains negative. Thus, the two possible candidates for the optimal A&E network configuration are either a merged pooled system (which would make the pooling strategy more attractive) or a pooled system for A&E 1 and a split system for A&E 2 (which would not change the optimal configuration).

(2) Assume $\delta^1 < 0$ and $\delta^2 < 0$. Recall that

$$\delta = \frac{1}{\lambda_s^1 + \lambda_s^2} \left[\left(\frac{c_g}{c_s} + p_s^1 - 1 \right) \lambda_s^1 + \left(\frac{c_g}{c_s} + p_s^2 - 1 \right) \lambda_s^2 - \left(\frac{1}{\kappa T_g} - \frac{1}{\kappa T_s} \right) \ln \left(\frac{1}{\alpha} \right) \right].$$

Thus, the fact that $\frac{c_g}{c_s} + p_s^1 - 1 < 0$ and $\delta^2 < 0$ shows that $\delta < 0$. Therefore, the optimal configuration for the two A&Es is to have a merged pooled system. When λ_s^1 is increased, we still have $\delta^1 < 0$ and $\delta^2 < 0$ and $\delta < 0$, which implies that the optimal configuration remains a merged pooled system.

(c) We prove the result by looking at all cases in Table 3. Because of the symmetry property, we only need to prove the result for either p_s^1 or p_s^2.

If λ_s^1 is increased, then δ^1 and γ^1 are increased, β^1 is decreased and δ^2, β^2 and γ^2 remain unchanged.

(1) Assume $\delta^1 \geq 0$ and $\delta^2 \geq 0$. If we increase λ_s^1, then δ^1 is increased. Therefore, we still have $\delta^1 \geq 0$ and $\delta^2 \geq 0$ and the optimal network configuration remains a merged split system.

(2) Assume $\delta^1 < 0$ and $\delta^2 \geq 0$. If we increase λ_s^1, then δ^1 is increased. Furthermore, if δ^1 is switched to positive after an increase in λ_s^1, then the new optimal network configuration becomes a merged split system, which would make the splitting strategy more attractive. Thus, we assume that δ^1 remains negative after an increase in λ_s^1.

If $\delta < 0$ and $\gamma^2 < 0$, then the optimal configuration for the A&E network is a merged pooled system. Thus, with an increase in λ_s^1, the new optimal network configuration would make the splitting strategy more attractive.

If $\delta < 0$ and $\gamma^2 \geq 0$, then the optimal configuration for the A&E network is to have a pooled system for A&E 1 and a split system for A&E 2. If we increase λ_s^1, then we have either $\delta < 0$ and $\gamma^2 \geq 0$ or $\delta \geq 0$ and $\gamma^2 \geq 0$. For the former, the new optimal network configuration remains unchanged, and for the latter, the new optimal network configuration becomes a merged system. Thus, in both cases, an increase in λ_s^1 would make the splitting strategy more attractive.

If $\delta \geq 0$ and $\beta^1 \geq 0$, then the optimal configuration for the A&E network is to have a merged split system. Note that

$$\delta = \frac{1}{\lambda_s^1 + \lambda_s^2}\left[\left(\frac{c_g}{c_s} + p_s^1 - 1\right)\lambda_s^1 + \left(\frac{c_g}{c_s} + p_s^2 - 1\right)\lambda_s^2 - \left(\frac{1}{\kappa T_g} - \frac{1}{\kappa T_s}\right)\ln\left(\frac{1}{\alpha}\right)\right].$$

Hence, if we increase λ_s^1, then δ remains non-negative and β^1 remains non-negative. Therefore, if we increase λ_s^1, then the new optimal network configuration remains a merged split system.

Note that the last case, where $\delta \geq 0$ and $\beta^1 < 0$, does not occur because $\frac{c_g}{c_s} + p_s^1 - 1 \geq 0$.

(3) Assume $\delta^1 \geq 0$ and $\delta^2 < 0$. An increase in λ_s^1 would still make $\delta^1 \geq 0$ and $\delta^2 < 0$.

If $\delta < 0$ and $\gamma^1 < 0$, then the optimal configuration for the A&E network is a merged pooled system. Thus, with an increase in λ_s^1, the new optimal network configuration would make the splitting strategy more attractive.

If $\delta \geq 0$ and $\beta^2 \geq 0$, then the optimal configuration for the A&E network is to have a merged split system. Note that

$$\delta = \frac{1}{\lambda_s^1 + \lambda_s^2}\left[\left(\frac{c_g}{c_s} + p_s^1 - 1\right)\lambda_s^1 + \left(\frac{c_g}{c_s} + p_s^2 - 1\right)\lambda_s^2 - \left(\frac{1}{\kappa T_g} - \frac{1}{\kappa T_s}\right)\ln\left(\frac{1}{\alpha}\right)\right].$$

If we increase λ_s^1, then we still have $\delta \geq 0$ and $\gamma^2 \geq 0$ and the new optimal configuration for the A&E network remains a merged split system.

If $\delta \geq 0$ and $\beta^2 < 0$, then the optimal network configuration is to have a split system for A&E 1 and a pooled system for A&E 2. Once again, if we increase λ_s^1, then we still have $\delta \geq 0$ and $\beta^2 < 0$. Thus, the new optimal configuration for the A&E network with an increase in λ_s^1 would be the same as the original optimal configuration for the A&E network.

If $\delta < 0$ and $\gamma^1 \geq 0$, then the optimal network configuration for the A&E network is to have a split system for A&E 1 and a pooled system for A&E 2. Then, an increase in λ_s^1 would not lead to $\gamma^1 < 0$. Thus, with

an increase in λ_s^1, the new optimal network configuration is either to have a split system for A&E 1 and a pooled system for A&E 2 (which would retain the same configuration) or to have a merged split system (which would make the splitting strategy more attractive).

(4) Assume $\delta^1 < 0$ and $\delta^2 < 0$. Recall that

$$\delta = \frac{1}{\lambda_s^1 + \lambda_s^2}\left[\left(\frac{c_g}{c_s} + p_s^1 - 1\right)\lambda_s^1 + \left(\frac{c_g}{c_s} + p_s^2 - 1\right)\lambda_s^2 - \left(\frac{1}{\kappa T_g} - \frac{1}{\kappa T_s}\right)\ln\left(\frac{1}{\alpha}\right)\right].$$

If $\delta \geq 0$, then an increase in λ_s^1 would still imply that $\delta \geq 0$. If an increase in λ_s^1 still keeps $\delta^1 < 0$, then the optimal network configuration before and after λ_s^1 is increased would be a merged split system. If an increase in λ_s^1 makes $\delta^1 \geq 0$, then δ remains non-negative and β^1 becomes non-negative because $\beta^1 \geq \delta^1$. Therefore, with an increase in λ_s^1, the new optimal network configuration would be a merged split system, which still makes the splitting strategy attractive.

If $\delta < 0$, then the optimal network configuration is a merged pooled system. Thus, any change in the new optimal network configuration after an increase in λ_s^1 makes splitting more attractive. □

Proof of Proposition 6

(a) We prove the result by looking at all cases in Table 3. Because of the symmetry property, we only need to prove the result for either p_s^1 or p_s^2. If p_s^1 is increased, then δ, δ^1, β^1 and γ^1 are increased but δ^2, β^2, and γ^2 remain unchanged. When $\delta^1 \geq 0$ and $\delta^2 \geq 0$, it is optimal to have a merged split system. When p_s^1 is increased, it still holds that $\delta^1 \geq 0$ and $\delta^2 \geq 0$ and with the new value for p_s^1, it is optimal to have a merged split system.

Assume $\delta^1 < 0$ and $\delta^2 \geq 0$. If $\delta < 0$ and $\gamma^2 < 0$, then it is optimal to have a merged pooled system. When p_s^1 is increased, the network configuration can be [P] or (P,S) or (S,P) or [S]. Thus, either the system configuration does not change or the splitting strategy becomes more attractive. If $\delta \geq 0$ and $\beta^1 \geq 0$, then it is optimal to have a merged split system. When p_s^1 is increased, it still holds that $\delta \geq 0$ and $\beta^1 \geq 0$ and with the new value for p_s^1, it is optimal to have a merged split system. If $\delta \geq 0$ and $\beta^1 < 0$, then it is optimal to have a separate (P, S) system. When p_s^1 is increased, both δ and β^1 are increasing. Thus, with the new value for p_s^1, it is optimal to have a separate (P, S) system or a merged split system. If $\delta < 0$ and $\gamma^2 \geq 0$, then it is optimal to have a separate (P, S) system. When p_s^1 is increased, both δ and β^1 are increasing but γ^2 does not change. Thus, with the new value for p_s^1, it is optimal to have a separate (P, S) system or a merged split system.

Assume $\delta^1 \geq 0$ and $\delta^2 < 0$. If $\delta < 0$ and $\gamma^1 < 0$, then it is optimal to have a merged pooled system. When p_s^1 is increased, the optimal network configuration can be [P] or (P,S) or (S,P) or [S]. Thus, either the system configuration does not change or the splitting strategy becomes more attractive. If $\delta \geq 0$ and $\beta^2 \geq 0$, then it is optimal to have a merged split system. When p_s^1 is increased, it still holds that $\delta \geq 0$ and $\beta^2 \geq 0$ and with the new value for p_s^1, it is optimal to have a merged split system. If $\delta \geq 0$ and $\beta^2 < 0$, then it is optimal to have a separate (P, S) system. When p_s^1 is increased, both δ and β^2 are increasing. Thus, with the new value for p_s^1, it is optimal to have a separate (P, S) system or a merged split system, which would make the splitting strategy more attractive. If $\delta < 0$ and $\gamma^1 \geq 0$, then it is optimal to have a separate (P, S) system. When p_s^1 is increased, both δ and γ^1 are increasing but γ^2 does not change. If with a new value for p_s^1, $\delta < 0$ and $\gamma^1 \geq 0$, then the optimal network configuration remains unchanged. If with a new value for p_s^1, $\delta \geq 0$, then the new optimal network configuration is either unchanged when $\beta^2 < 0$ or to have a merged split system when $\beta^2 \geq 0$. Therefore, with the new value for p_s^1, it is optimal to have a separate (P, S) system or a merged split system.

Assume $\delta^1 < 0$ and $\delta^2 < 0$. If $\delta \geq 0$, then it is optimal to have a merged split system. When p_s^1 is increased, δ^1 and δ are increasing, but δ^2 remains unchanged. If with the new value for p_s^1, $\delta^1 < 0$, then it is optimal to have a merged split system. If with the new value for p_s^1, $\delta^1 \geq 0$, then we have shown that $\delta \geq 0$ and will show that $\beta^2 \geq 0$. Recall that

$$\delta = \frac{1}{\lambda_s^1 + \lambda_s^2}\left[\left(\frac{c_g}{c_s} + p_s^1 - 1\right)\lambda_s^1 + \left(\frac{c_g}{c_s} + p_s^2 - 1\right)\lambda_s^2 - \left(\frac{1}{\kappa T_g} - \frac{1}{\kappa T_s}\right)\ln\left(\frac{1}{\alpha}\right)\right].$$

Thus, $\delta \geq 0$ and $\delta^1 < 0$ implies that $\frac{c_g}{c_s} + p_s^2 - 1 \geq 0$, which in turn shows that $\beta^2 \geq 0$. Therefore, with an increase in p_s^1 such that $\delta^1 \geq 0$, the optimal network configuration is a merged split system. This shows that the optimal system remains unchanged when we increase $p^1 - s$.

If $\delta < 0$, then it is optimal to have a merged pooled system. Consequently, with an increase in p_s^1, the new optimal network configuration would make the splitting strategy more attractive. For the remaining cases (b)–(d), we have: (b) When both p_s^1 and p_s^2 are sufficiently close to 1, we have $\delta \geq 0$, $\beta^1 \geq 0$ and $\beta^2 \geq 0$. Therefore, the results in Table 3 show that it is optimal to have a merged split system. (c) When both p_s^1 and p_s^2 are sufficiently close to 0, we have $\delta < 0$, $\beta^1 < 0$ and $\beta^2 < 0$. Therefore, the results in Table 3 show that it is optimal to have a merged pooled system. (d) When p_s^1 is sufficiently close to 1 but p_s^2 is sufficiently close to 0, we have $\beta^1 \geq 0$, $\beta^2 < 0$, $\gamma^1 \geq 0$, and $\beta^2 < 0$. Therefore, the results in Table 3 show that it is optimal to

have a split system for A&E 1 and a pooled system for A&E 2. The same can be said for the case where p_s^1 is sufficiently close to 0 but p_s^2 is sufficiently close to 1. □